

Is the university–industrial complex out of control?

Links between academia and industry are of increasing concern to academics and to society at large. The sectors involved should review their policies in order to sustain universities' public accountability.

In recent years, many universities have thrived on their relationships with industry. In the sciences, academics and their institutions have benefited from contracts, donations and sponsorship, and from the entrepreneurialism of faculty members — a trait now encouraged by government in most scientifically developed countries. Nowhere is this more so than on the west coast of the United States. One third of all the world's biotechnology companies were founded by faculty members of the University of California.

The benefits to researchers and universities are many. They include access to industry facilities and databases, financial support for research that can help the university as well as the company, opportunities for academics to tap into the market's expertise, and the longer-term benefits of experience and contacts.

The downside — for academia at least — of these benefits is becoming increasingly clear. Some of the problems arise in the scientific literature. Recent publications in biomedical journals indicate that researchers sponsored by companies are biased in favour of reporting positive experimental results relating to company products. Undeclared conflicts of interest have occasionally undermined trust in published research and reviews. Other difficulties arise if companies try to restrict academic freedoms or institutionalize industry's influence. For example, in at least one contract offered to a prestigious university by a multinational chemical company, the company took ownership and rights over all data deposited in their database. The academic involved found the company inflexible on this issue and so withdrew.

One controversial example of industrial influence is the deal between the University of California at Berkeley and Novartis. The company is paying about \$5 million per year for plant research and providing access to its databases; in return, it gains a seat in university and departmental research committees and restricts academics' freedom to discuss the benefits of the deal. Some students and academics are protesting against the agreement, claiming it undermines Berkeley's status as a publicly funded and publicly accountable institution.

Anti-corporatism vs anti-liberalism

Protests against over-intimate industrial ties are not restricted to Berkeley. Young biologists at a meeting on genetics and society organized by the European Commission lamented their universities' increasing dependence on industrial funding, and one patients' group attacked another for its links to a pharmaceutical company.

Such protests can be seen as the more informed end of the grouping known as the post-Seattle movement. This movement represents an active unease within society: that global corporations have too much unaccountable influence on institutions — including universities — that are meant to act in the public interest. The Novartis–Berkeley deal can all too easily be portrayed as an institution undermining both its motivation and trustworthiness to provide an independent and impartial view of one of the most

contentious technologies of our time — genetically modified crops. Although it is unusual in its scope, other universities will recognize the issues it raises.

The chancellor of Berkeley, Robert Berdahl, although standing by the Novartis deal, has expressed his own unease at what he calls the privatization of the public universities. In a recent speech (see www.chance.berkeley.edu/cio/chancellor/sp/privatization.htm), he highlighted two forces that are changing the environment for public universities in the United States: declining public funds, and a systematic, successful and entirely legitimate campaign to develop an anti-liberal, right-wing agenda in universities. These two forces, he implies, are stimulating the growth of the university–industrial complex. And the dangers? Loss of cohesion at the university due to salary differentials and market forces; a downgrading of humanities — while the technology-related ethical and social issues that they should address are burgeoning — and a loss of academic objectivity and consideration of the wider world.

New year's resolutions

So what resolutions are required to maintain public trust in higher education and publicly funded research?

- **Vigilance and the ability and determination to speak out.** Caught between anti-corporatism and anti-liberalism, academics have to stand up for themselves, with the protection of university constitutions and hierarchies.
- **Transparency over conflicts of commitment.** There is, for example, a need to prevent the valuable findings of government-funded students and postdocs being inappropriately funnelled into commercial ventures. This could prevent timely publication and lead to commercial ownership of a discovery that was made using public money. There should be limits on, and disclosure of, involvement of faculty members and others in companies.
- **Transparency over conflicts of interest.** Recent surveys in biomedical journals reveal the pervasive (though highly variable) rules in US universities on disclosure of financial interests. (This journal intends to introduce a disclosure policy of its own in the near future.)
- **The issues need to be debated and dangers highlighted at a national level.** In this respect the imminent inquiry into university–industrial links by the US national academics' committee on science, engineering and public policy is timely.
- **Industry must help sustain public trust.** Some companies approach universities and the knowledge they produce in a way that can be best described as predatory. Occasionally this may yield short-term returns, but public distrust of research can have a powerful impact on regulation and on consumers.

And regional and national governments, although encouraging the university–industrial complex, must keep watch over its development. They must above all underpin the social value and accountability of public universities with strong financial support. ■

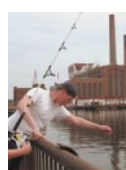




On the defensive
Bush administration
set to back US missile
shield
p122



Aiming high
Astronomers in
Canada seek
funding boost
p123



Mercury rising
US power plants face
prospect of mercury
emissions control
p124



Full house
Moderate named as
chair of Science
Committee
p125

Scepticism greets claims that uranium shells cause leukaemia

Quirin Schiermeier, Munich

This month's fierce political debate in Europe over the health hazards posed by ammunition made from depleted uranium is being met with broad scepticism from researchers. Instead, they suggest a number of potential candidates for the health problems being observed.

The argument centres on shells used in the 1999 Kosovo conflict, which were made from uranium-238, a mildly radioactive by-product from nuclear power production.

Italy, Belgium, Portugal and Sweden have each started separate investigations into the health effects of the ammunition following the discovery of six cases of leukaemia among Italian soldiers who served in Kosovo.

In response to mounting public concern, NATO says it will provide a full list of targets in Serbia and Kosovo where depleted uranium ammunition was used in attacks.

But radiation experts say that exposure to depleted uranium, which is used in ammunition on account of its very high density, would be unlikely to cause blood-based diseases such as leukaemia. The alpha radiation it emits is less intense than that from the naturally occurring element, they say.

"If anything, we would expect depleted uranium to cause lung or bone cancer, rather than leukaemia," says Manfred Paschke, head of radiation protection at a German government nuclear research laboratory in Jülich. Uranium intake is known to cause kidney damage, but the scientific literature shows no link with leukaemia.

Researchers say that the investigation of leukaemia in Kosovo veterans should not be confined to uranium. Enforcement of health and safety rules in the military during wartime are "notoriously lax," says Paschke. "The cancer risk from misuse of lubricants, solvents and disinfectants is always high," he adds.

The military, environmental and health implications of uranium-238 have been investigated since at least 1974. The US Department of Defense's latest report on the matter, published last month in response to concerns over exposure of troops during the Gulf War, finds that: "The evidence to date does not support claims that depleted uranium caused or is



NATO's legacy? Portuguese soldiers measure radiation levels near Klina in Kosovo last week.

causing Gulf War veterans' illnesses."

Scientists believe that this also applies to Kosovo. According to press reports, only around eight tonnes of uranium-238 were released by US bombing of Serbian targets during the Kosovo war, against about 350 tonnes used in the Gulf conflict. "This is much

too little for a health-sensitive large-scale contamination of the region," says Christian Küppers, a nuclear radiation expert at the Öko Institute in Darmstadt, Germany, which does environmental health research.

♦ <http://www.gulflink.osd.mil/envexp.html>

♦ <http://www.llrc.org/durs.htm>

Dye dispute leans Amersham's way

Declan Butler

Rival manufacturers of automated DNA-sequencing machines have clashed in a California courtroom over the intellectual property rights to the fluorescent dye used in their equipment.

According to a pretrial ruling made last week by US federal judge Charles Breyer, the dye technology used by the California-based Applied Biosystems infringes a patent held by the British company Amersham Pharmacia Biotech. The case will come to full trial later this month.

The ruling is one of several in a tit-for-tat battle over various patents and claims on sequencing technology, which began when Amersham sued Applied Biosystems in 1997 over the launch of its BigDye product.

At stake are not only the market shares of Applied Biosystem's 3700 Prism DNA analyser and Amersham's MegaBACE system, but also a slice of the glory for the sequencing of the human genome.

In a written statement in response to the pretrial ruling, Applied Biosystems said that the company "continues to believe that we will prevail". But Sir William Castell, chairman of Amersham, describes the ruling as "very significant". He adds that the onus is now on Applied Biosystems to invalidate the pretrial ruling.

Amersham is suing Applied Biosystems for undisclosed financial damages. The trial is expected to end in February, with a 90-day period for appeals.

♦ <http://www.delphion.com/details?pn=US05688648>

Overseas investors plan six 'world class' universities in India

K. S. Jayaraman, New Delhi

Wealthy Indian entrepreneurs based in New York are planning to invest up to \$1 billion to establish six "world class" private research universities across India.

Although similar proposals have been resisted in the past as a potential threat to existing institutions, the new plan has been endorsed by Prime Minister Atal Behari Vajpayee. "We will actively facilitate this and other similar initiatives," he told the 88th Indian Science Congress in New Delhi last week.

The backers of the proposed Global Institutes of Science and Technology (GIST) say that "the need for science and technology talent will continue to rise in India as well as globally, and that India will be a major contributor of that talent".

GIST will be privately funded and managed with no financial input from India's government. It will be registered as an Indian company and, the backers say, will have a president accountable to a governing board made up of major sponsors, world-renowned scientists and chief executive officers of multinational corporations.

Each of its six campuses will have a yearly input of 2,000 students, and GIST plans to form affiliations with leading US research universities.

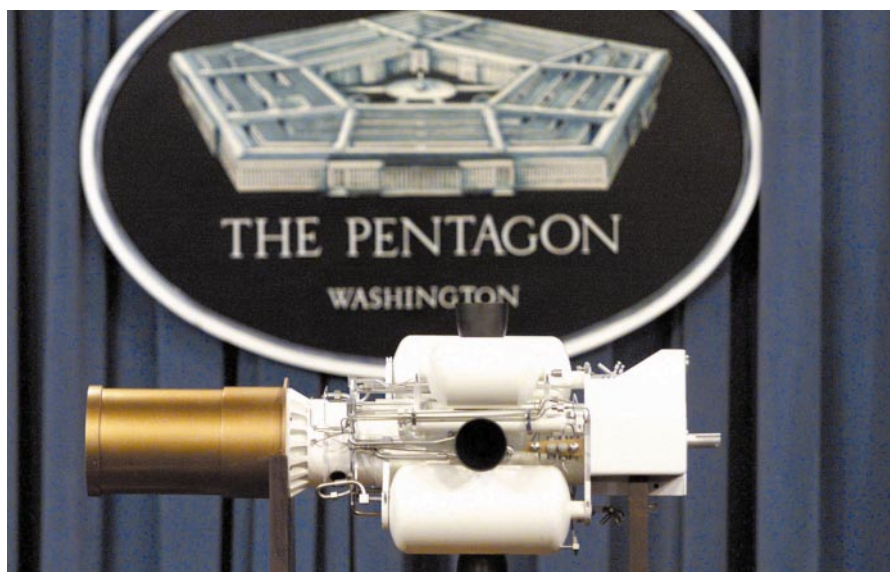
The proposal was approved last November by the prime minister's science advisory committee, but it has yet to be formally cleared by India's Ministry of Human Resources Development.

Vajpayee also said his government would double research funding by 2004, reform the structure of higher-educational institutions, and free scientific institutions from bureaucratic control. In return, he asked the scientists to transform India into a leading scientific nation in the twenty-first century. ■

► <http://isc2001.nic.in/vsisc2001>



Indian Inc.: the Indian Science Congress heard how private money will fund new universities.



In for the kill: a model 'kill device' from the missile defence system tested by the Pentagon last year.

Bush targets space-based missile defence system

Irwin Goodwin, Washington

George W. Bush's choice of defence secretary shows that his new administration aims to give the United States a missile defence shield — confirming the fears of some scientists and of US allies.

But some scientific groups say that Bush's choice of Donald Rumsfeld for the position could also herald a push by the United States to establish military supremacy in space.

Rumsfeld is a champion not only of missile defence, but also of US efforts to control space by developing the technology to protect satellites in orbit. Such initiatives could militarize space in the next two decades, a prospect that some defence experts have long urged and others have passionately opposed.

Rumsfeld, who was President Gerald Ford's White House chief of staff and then secretary of defence 25 years ago, chaired a bipartisan commission for Congress in 1998 that helped to build political support for missile defence by warning that Iran, Iraq and North Korea were closer than had been thought to being able to direct nuclear or biological warheads at the United States.

Now, another congressionally mandated commission headed by Rumsfeld is completing a report on the threats to US satellites, which are increasingly vital to military and civilian communications. The report, expected this month, will endorse "US control of space, including defending our own satellites and engaging those of any enemy," according to one member of the panel.

One sceptic is space-policy analyst John Pike, formerly at the Federation of American Scientists and now at Globalsecurity.org.

Pike says that members of Congress pushed for the satellite study because they believe the United States should devise anti-satellite weapons and consider establishing a space force as an arm of the Defense Department. He opposes the idea, arguing that it is "singularly misguided, when we are the only nation with satellites worth shooting at".

The National Missile Defense (NMD) planned by President Bill Clinton would have been land-based, with interceptor missiles ready for launching from Alaska by 2007. It calls for satellites that would track missiles carrying weapons of mass destruction.

Bush's defence system, by contrast, would use space weaponry. His vision is reminiscent, some observers say, of President Ronald Reagan's 1983 dream of an impenetrable Strategic Defense Initiative, dubbed Star Wars. Although the Pentagon has been working for years on lasers that might someday be mounted on aircraft or satellites, such weapons do not yet exist.

Clinton's proposed NMD is already undergoing painful reappraisal, after two out of three early tests failed. But the Pentagon still plans to attempt at least three missile interceptions this year, and then proceed until 21 tests have been successful.

The Union of Concerned Scientists and the Security Studies Program at the Massachusetts Institute of Technology have concluded that nations that could launch missiles at the United States could also develop countermeasures to defeat a limited defence system. The Cornell University physicist Hans Bethe gave Congress the same advice more than 30 years ago. ■

Canadian astronomers mount lobby effort

NASA

Steve Nadis

Tired of meagre grant support and alarmed at their possible exclusion from international projects, Canadian astronomers are mounting an aggressive lobbying campaign to win more financial support.

Canadian astronomy is rated third in the world in terms of citations per paper, despite, Canadian astronomers claim, the government spending seven times less per head than the United States.

But Canadian astronomers say they will be unable to maintain their position without a significant financial boost. "The rules of the game have changed," says University of Calgary astronomer Russ Taylor, president of the Canadian Astronomical Society. "Astronomy is becoming more international and the stakes are getting bigger." More resources are needed, he says, for his country to participate in the observatories of the future.

The Coalition for Canadian Astronomy, which is made up of academics and industrialists, has been formed to persuade the government to increase funding. "A campaign of this kind, which puts all astronomy under a single umbrella, is a first for us," says Ralph Pudritz, an astronomer at McMaster University, Ontario, who serves on the coalition's executive committee. "We're slowly catching on to the American way of doing things."

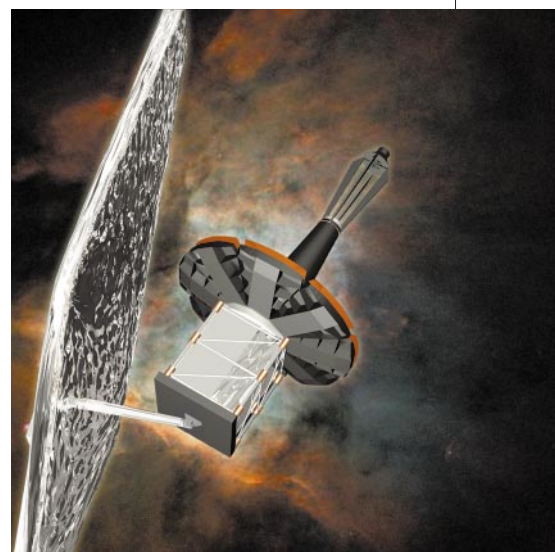
Pudritz chaired a national panel that authored a report, published last summer, which set priorities for Canadian astronomy over the next decade. The coalition wants to implement this plan. "We're making the case as an entire community, rather than vying for support for individual projects as we've done in the past," Pudritz says.

The coalition is calling for Can\$16.4 million (US\$10.9 million) per year over the next ten years for ground-based astronomy; about twice the current funding. The top priority will be involvement in the Atacama Large Millimeter Array (ALMA) in Chile, destined to become the world's most powerful radio telescope in millimetre and sub-millimetre wavelengths.

Last month, Canada's National Research Council and the US National Science Foundation reached an agreement to give Canadian astronomers access to ALMA and all major US radio observatories.

Researchers also aim to contribute to the designs of the Square Kilometre Array, a centimetre-wavelength radio telescope, and the Very Large Optical Array.

The coalition is seeking Can\$75 million over ten years to allow participation in NASA's Next Generation Space Telescope and another Can\$25 million for the European Space Agency's Planck mission



High hopes: Canadian astronomers hope to join the Next Generation Space Telescope project.

to study cosmic background radiation.

Taylor and other coalition members were in Ottawa this week, lobbying members of parliament. "This is make or break for us," says Pudritz. "If we're not part of the world observatories, recruitment efforts will suffer and we'll lose our edge within a decade."

Scientists unimpressed by Australia's funding plan

Peter Pockley, Sydney

Australian scientists have reacted ambivalently to the conservative government's plan to inject new funding into research.

A dozen programmes would be started or expanded under a plan released late last year by Robin Batterham, the government's chief scientist. He proposes adding A\$1.8 billion (US\$1 billion) to government spending over five years, most of it for university-based research.

The largest expansion would be a doubling of grants through the Australian Research Council, costing A\$660 million, a

A\$150 million expansion of the Cooperative Research Centres, and A\$400 million of extra funds for major new research facilities.

The plan would try to leverage more industrial research, which has been in decline in Australia. Batterham says the country's competitive position has been "reasonable" in the past, but admits that it must now be improved. "Finland and Ireland over ten years have more than doubled their business investment" in research and development, he says. "We haven't ... the case presented is too powerful to ignore."

The government is expected to respond to Batterham's proposal in a statement on innovation to be issued early this year. Although universities and research bodies are urging the government to implement Batterham's plan in full, they are separately calling for larger boosts in general support to restore Australia's investment to previous levels (see chart, left).

Barry Osmond, a plant biologist who is leaving Australia to head Columbia University's Biosphere 2 facility in Arizona (see *Nature* 408, 396; 2000), is unimpressed by the plan. Australian science "had a



Batterham believes business must invest more in research.

tremendous reputation", he laments.

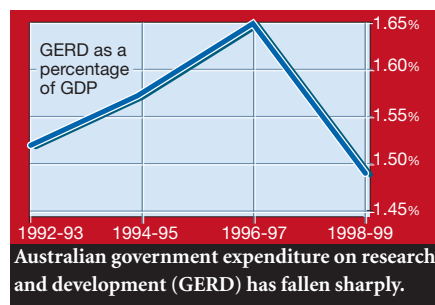
"Government or the bureaucracy have just let science run down to the point where it's on the verge of irredeemable damage. The current proposals are peanuts. It needs billions of dollars per year. Compared to our competitors in this

field, we're doing nothing," Osmond says.

An analysis by Australia's top eight research universities says it would take A\$9.4 billion over five years to raise the country's research spending to the developed world's average of just over 2% of gross domestic product (GDP).

Since the middle of last year, the plight of Australian science has attracted considerable media coverage, putting pressure on the government to support it as it enters an election year.

AUSTRALIA BUREAU OF STATISTICS



PETER POCKLEY

US to take temperature of mercury threat

Mark Schroppe

In the United States alone this year, some 60,000 babies may be born with neurological damage caused by mercury poisoning of their mothers.

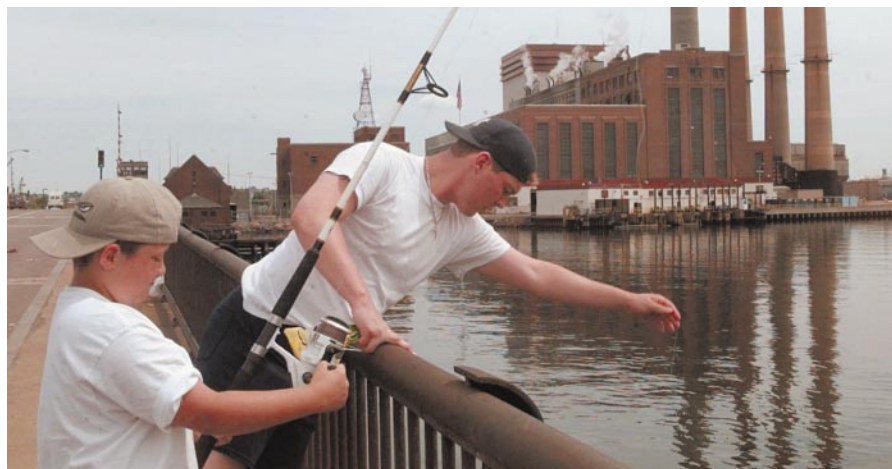
Scientists believe that much of the mercury enters the food chain through seafood whose habitat has been polluted by emissions from power stations burning mercury-rich coal. But it would be expensive — and therefore controversial — to regulate such emissions.

However, as one of several parting shots, President Bill Clinton's Environmental Protection Agency (EPA) last month laid out plans to implement such regulation by 2004 — and a research strategy to back it up.

Jonathan Herrmann, assistant director of the EPA's National Risk Management Research Laboratory in Cincinnati, Ohio, and a co-leader of the team that devised the strategy, says that the nature of the regulations will depend on the research. The plan is to explore the feasibility of different approaches to mercury removal, the distribution of emissions, and their effects on human health.

Some mercury is already removed from power-plant emissions along with other pollutants. The research will investigate by how much mercury emissions can be reduced using existing and emerging technologies, the relative cost of reductions, how much of the mercury in the environment can be attributed to power plants, and the probable effect of regulation on environmental mercury levels.

The answers will help the agency to determine not only how much mercury the power plants should be required to remove, but how much removal is economically reasonable. Currently, the EPA thinks regulation will cost



Throw it back: mercury emitted by power plants enters fish, and then the people that eat them.

between one billion and two billion dollars to implement, whereas industry estimates range from two billion to six billion dollars.

"It seems like a well-balanced strategy and I support wholeheartedly the work that they're doing," says Leonard Levin, a manager at the Electric Power Research Institute, based in Palo Alto, California. Levin, who was one of the strategy's outside reviewers, says it should provide a framework for establishing mercury regulations.

Mercury is released from power plants and waste incinerators in an inorganic form. But in water, microbes convert it into methylmercury, which accumulates in fish and shellfish. When the fish are eaten, methylmercury accumulates in the brain, where it reverts back to an inorganic form.

A National Academy of Science (NAS) study last year estimated that the mercury

problem was substantial enough to influence the total number of children who struggle in school and require special help.

The NAS report, which contains the estimate that 60,000 mothers could be affected each year, also recommended research to understand the extent of mercury's effects on humans. For instance, there is some evidence that mercury causes cardiovascular problems in children.

Thomas Burke, an epidemiologist at Johns Hopkins University in Baltimore, and a lead author of the NAS study, says that, despite gaps in understanding, increased regulation is warranted.

"There is nothing to indicate that mercury risks are less than we thought," he says. "I think it is very appropriate to move forward aggressively to limit mercury in every way possible."

Hunt for Earth-like planets edges towards launch pad

Tony Reichhardt, Washington

A space telescope could search for Earth-sized planets orbiting other stars is among three mission concepts picked by NASA last week to compete for a launch slot in 2005 or 2006.

The Kepler mission and two other proposals — Dawn to orbit the asteroids Vesta and Ceres, and INSIDE Jupiter to investigate the giant planet's interior by studying its magnetic and gravitational fields — were chosen as finalists from 26 ideas submitted to the agency's Discovery programme in August. Each team will receive \$450,000 to define its concept further, and a winner will be selected late this year. All three missions are priced at just under \$300 million.

The choice of Kepler as a finalist is a

vindication of sorts for principal investigator William Borucki of NASA's Ames Research Center in California. He has long proposed a novel and controversial method of searching for Earth-sized planets.

Placed into orbit around the Sun, and equipped with a 0.95-metre telescope and ultra-sensitive charge-coupled device (CCD) detectors, Kepler would fix its gaze on 100,000 stars simultaneously, looking for extremely small dips in their light output. An Earth-sized planet passing in front of a Sun-like star would eclipse its light by only a few parts in 100,000.

But that would be enough to betray the planet's existence, says Borucki, if the pattern is regular. If Earth-sized planets are common, Kepler could detect more than 500 during its four-year lifetime.

Borucki has had trouble convincing NASA review committees that a spacecraft can achieve the sensitivity needed to register such minute changes in light. But he has since successfully tested his system in the laboratory, even when introducing the types of 'noise' that might be expected in space.

Planet hunters have found dozens of Jupiter-sized objects, but nothing like Earth. Borucki and his colleagues claim that their technique is the only one that can find Earth-sized planets in the zone near a star where water would remain liquid at the surface.

But Kepler would not be the first such instrument to fly. The French-led Corot mission is scheduled to launch in late 2004, and will use a CCD array to look for transiting extrasolar planets.

► <http://discovery.nasa.gov/news.html>

Moderate takes key science role in Congress

Colin Macilwain, Washington

A moderate Republican with a keen interest in environmental and energy issues has been named as chair the Science Committee in the US House of Representatives.

The selection of Sherwood Boehlert (Republican, New York) to head the panel is good news for the scientific community at a time when it awaits, with uncertainty and even some trepidation, news of the science administrators in President George W. Bush's administration.

Sherry Boehlert is regarded as smart, sympathetic to science and — critically for the science panel — effective at getting results in Congress. Since 1995, when Republicans won control of the House, Boehlert has skilfully exploited his position

as a moderate Republican to foster influence, especially on environmental policy.

"I'm looking forward to an exciting couple of years," Boehlert said in an interview the day after his selection. "Science policy is of critical importance to our future as a nation." He pledged to work for additional funding for the National Science Foundation (NSF), the National Institutes of Standards and Technology (NIST), and other agencies under the committee's jurisdiction.

"We've got a wide-open portfolio, and it would be presumptuous of me to lay out an agenda at this stage. But one thing we will need to do is to help to develop a comprehensive, coherent energy policy."

The Science Committee oversees most



Man in the middle: Boehlert has used his moderation to build influence on Capitol Hill.

US research that is not medical or military, including the space agency NASA and civilian research at the Department of Energy as well as the NSF and NIST. As the only committee in either house of Congress devoted to science, it has traditionally served as an important forum for scientific issues.

But under its past two chairs, Robert Walker (Republican, Pennsylvania) and James Sensenbrenner (Republican, Wisconsin), science lobbyists have seen the committee as partisan and insufficiently committed to the cause of the research agencies under its jurisdiction.

Sensenbrenner alienated the scientific community by failing to back a bill that would have authorized the doubling of research spending, and by pursuing what many saw as a political vendetta against the Advanced Neutron Spallation Source project at the Oak Ridge National Laboratory in Tennessee.

Boehlert's arrival is likely to result in a resumption of the wide-ranging discussion of science that the committee conducted under its most recent Democratic chairman, the late and much-lamented George Brown. "We are going to move in the direction of committee staff having solid scientific backgrounds," says Boehlert.

As for fears that his moderate reputation will weaken the committee — Boehlert is the only one of 13 Republican House committee chairs named last week who is not seen as a conservative — he views it as a strength.

"I don't sense any isolation," he says. "I'm where the overwhelming majority of the American people are — in the middle ground."

Boehlert also hopes that his friendship with Senator John McCain (Republican, Arizona), which dates from their arrival in the House together in 1982, will help the Science Committee to get Senate attention for its legislative ideas. McCain is chair of the Senate's Commerce, Science and Transportation Committee.

Biologists crusade for evolution

Rex Dalton, Chicago

A newly energized effort to promote the teaching of evolution is underway across the United States, spurred by scientists who think the time is ripe to combat the creationist movement.

At its annual meeting last week, the Society for Integrative and Comparative Biology (SICB) became one of about a dozen organizations that have recently adopted policies to boost the teaching of evolution at all levels of education.

The drive follows the National Conference on the Teaching of Evolution, held at the University of California at Berkeley, where representatives of nearly 50 organizations met in October to plot a national consensus on promoting the teaching of evolution.

Individual societies — ranging from large organizations like the American Institute of Biological Sciences to small groups like the Association of Systematics Collections — are being urged to encourage

their members to become actively involved in teaching evolution.

Berkeley officials hope to receive \$500,000 from the US National Science Foundation to set up a website showing ways to teach evolution. The three-year project will also study its own effectiveness.

"We want to create one-stop shopping for evolution teaching for instructors, from kindergarten through the university level," said Judy Scotchmoor, coordinator of the evolution programme at Berkeley's Museum of Paleontology.

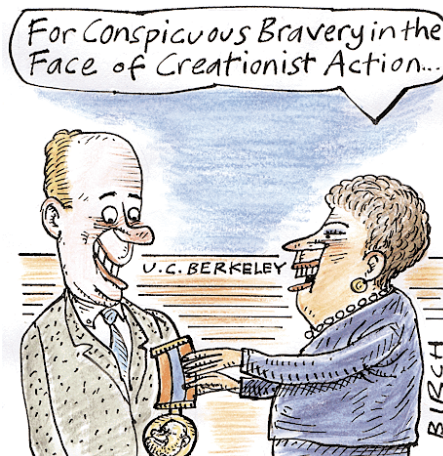
At the annual meeting of the SICB, whose 2,100 members are specialists in developmental biology, a discussion was held on how to ensure the widespread teaching of evolution.

Kevin Padian, a palaeontologist at Berkeley, told the meeting it was important that the chairs of university departments recognized the importance of allowing faculty to assist in grassroots efforts to further evolution teaching.

Professors should receive academic credit for engaging in the battle, said Padian, adding that "it is just as important as putting a paper in *Nature*, *Science* or the *American Zoologist*."

Ronald Edwards, an evolutionary biologist at DePaul University in Chicago, suggested that professors should not be confrontational or arrogant with students who have been taught creationism in their churches or homes.

"College students are willing to take on new ideas, but they won't if there is a hint of antagonism," said Edwards, who spoke of his experiences at Valdosta State University in Georgia, where many students believe in creationism.



Guidelines reject criticism of policy on gene patenting

Washington The US Patent and Trademark Office has released new guidelines on the patenting of genes and gene sequences that will clarify its handling of applications.

The new guidelines are broadly similar to an interim version that was released for comment in December 1999 (*Nature* 403, 3; 2000). In particular, they accept that gene patents lie within the current laws governing what can be patented, and reject broad criticism of the current gene-patenting policy.

However, the final version of the guidelines provides greater clarification of the circumstances in which a claim for the usefulness — or 'utility' — of a newly discovered gene sequence can be based on its similarity to a previously known sequence of known function.

This issue had proved one of the most contentious elements of the proposed revisions, with a range of organizations, including the National Institutes of Health and the medical colleges, arguing that the original guidelines were too loose to establish genuine inventiveness.

The new guidelines are intended to cover

areas of emerging technologies "where uses for new materials that have not been fully characterized are not readily apparent".

♦ <http://www.uspto.gov>

CNN founder funds centre to counter nuclear threat

Washington Ted Turner, the founder of global news corporation CNN, has announced that he will provide \$250 million over five years to support a new institution aimed at reducing the global danger of nuclear weapons.

Charles Curtis, a former deputy energy secretary, has been named as president of the organization, which is provisionally called the Global Threat Reduction Institute.

Turner had the idea of the institute in discussions with Sam Nunn, the former Democrat senator for Georgia. Nunn had previously worked with Curtis on a study of the nuclear threat. The institute will try to increase public awareness of the nuclear threat, and will support research into arms control and antiproliferation strategies.

X-ray work leading to Indian satellite launch

New Delhi Indian scientists have received the go-ahead to develop and build instruments for an astronomy satellite. But the Indian



Reach for the stars: the Polar Satellite Launch Vehicle should take the new satellite into orbit.

Space Research Organisation (ISRO), which is funding the project, says the satellite's launch will depend on progress over the next 18 months in developing three X-ray instruments for the mission, including an X-ray space telescope.

"We have to develop the prototype instruments in this period and show that we can successfully make them in India," says

P. C. Agrawal of the Tata Institute of Fundamental Research in Mumbai. He says the proposed payload is much more complex than the one his team built for an Indian remote-sensing satellite launched in 1996.

ISRO spokesman S. Krishnamurthy says that the satellite has been proposed for studying a variety of cosmic sources, ranging from nearby stars to distant extragalactic objects such as quasars. The payload will be built at the Tata institute and the satellite will

be launched by ISRO's Polar Satellite Launch Vehicle. He says the agency is aiming for a launch "in about five years".

Canada 'faces 70% shortfall in specialist graduates'

Montreal Canada faces a 70% shortfall between the number of graduates in certain disciplines and industry's requirement for them, according to a survey by campaigners.

Since 1995, the number of graduates taking masters' degrees in Canada has fallen by 4% and the number of PhDs by 35%. The problem is particularly acute in engineering and computer science, says Doug Barber, who chairs the recently formed campaigning group eMPOWER — named for its focus on microelectronics, photonics, optoelectronics and wireless and radio engineering. Barber, the former president of electronics company Gennum, has called for a tripling of faculty members in all those disciplines.

The group is demanding Can\$500 million in federal funding for university education.

Estonians give green light to gene bank

Tallinn The Estonian parliament last week approved, by a large majority, the Human Gene Research Act. This legislation will

regulate the creation, maintenance and use of a national gene bank in Estonia.

In accordance with normal ethical constraints, data in the gene bank will be correlated with health data and will be used to help locate genes that are associated with complex diseases.

The act also gives the go-ahead for the creation of a foundation to fund the project, called the Estonian Genome Project, to the tune of US\$150 million during the next five years. The government will provide 20% of costs, but the rest must be found from private investors.

Catalonia aims to entice expat researchers home

Barcelona The Spanish autonomous community of Catalonia has launched a new research institution aimed at enticing Catalan researchers working elsewhere to return home.

The Catalan Institution of Research and Advanced Studies will be supported by the Catalan government research department and the Catalan Research Foundation. The institution hopes to hire 25 researchers a year.

Andreu Mas-Colell, head of the research department, hopes that, in combination with other initiatives, Catalonia will be able to attract back 100 researchers each year.

Sulston knighted for human genome work

London John Sulston, the former director of the Sanger Centre near Cambridge and a leading figure in the international human genome project, has received a knighthood in the New Year Honours list.

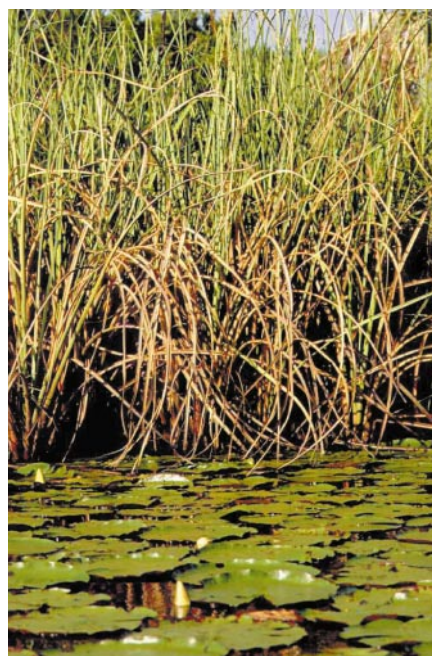
Sulston says he hopes his honour will draw attention to the enormous group working on the project at the Sanger and elsewhere. In the next few years, he plans to continue working on the genome project and to study nematode functional genomics with his long-time collaborator, Alan Coulson.

Other scientists who were knighted include biotechnology entrepreneur Christopher Evans, who started companies including Enzymatix and Merlin Ventures, and Chris Llewellyn Smith, the former director-general of CERN.



Sir John Sulston: an honour for genome project.

PA/VEA



Save our swamp

Florida's Everglades are the focus for the largest ecological restoration project yet attempted. But critics claim the present plan will fail to halt the wetlands' decline. Mark Schrope investigates.

In the 1920s, the southwards march of development in Florida ran up against the vast wetlands of the Everglades. Over the decades that followed, the construction of an elaborate network of levees, canals and dams allowed a boom in agriculture and population. Today, about half of the original wetlands are gone, and a large volume of water that once flowed through the Everglades is now diverted into the Atlantic and the Gulf of Mexico. Numerous plants and animals, such as the American crocodile, *Crocodylus acutus* are in dramatic decline, and wading bird pop-

ulations have dropped to about one tenth of their former strength.

In December, as the world's media focused its attention on Florida's disputed election, the outgoing US president, Bill Clinton, signed a bill authorizing the first tranche of spending on a plan to rescue the Everglades. The Comprehensive Everglades Restoration Plan is billed as the largest ecological restoration project ever attempted. Its total cost, spread over two decades, is expected to reach \$8 billion, with the state of Florida and the federal government sharing the tab.

The plan is largely the work of the South Florida Water Management District

(SFWMD) — a state agency based in West Palm Beach — and the US Army Corps of Engineers. It calls for land to be bought and flooded to allow better storage and release of water through the ecosystem. It also aims to remove some barriers to flow, and build a water control network of pumps and water treatment and storage areas.

But among hydrologists and wetland ecologists, there is a growing consensus that the plan is seriously flawed. Scientific critics argue that, in focusing primarily on restoring historical depths of water, too little attention has been paid to patterns of water flow. They fear that the plan will allow substantial further degradation of the Everglades. And some marine ecologists are concerned that the plan could also threaten Florida Bay, further damaging its already fragile health.

Everyone agrees that there are no simple answers. Returning the Everglades to their natural state is not an option. The project's goal is instead to restore flow as much as possible while also considering the needs of local residents and farms.

Breach the barricades

The present plan would increase water movement through existing barriers, for instance by raising small sections of highways on to bridges, or cutting breaks into levees. At the same time, it would imitate aspects of historical flows using the water control network. The critics' central concern, based on recent studies of the Everglades' hydrology, topography and palaeoecology, is that too few barriers are being removed. "If you're going to restore that ecosystem," says Stuart Pimm, an ecologist at the University of Tennessee in Knoxville, "you have to let this free-flowing system be a free-flowing system."

These concerns stem in part from research by Christopher McVoy, a hydrologist with the SFWMD. McVoy has pieced together a picture of what the Everglades



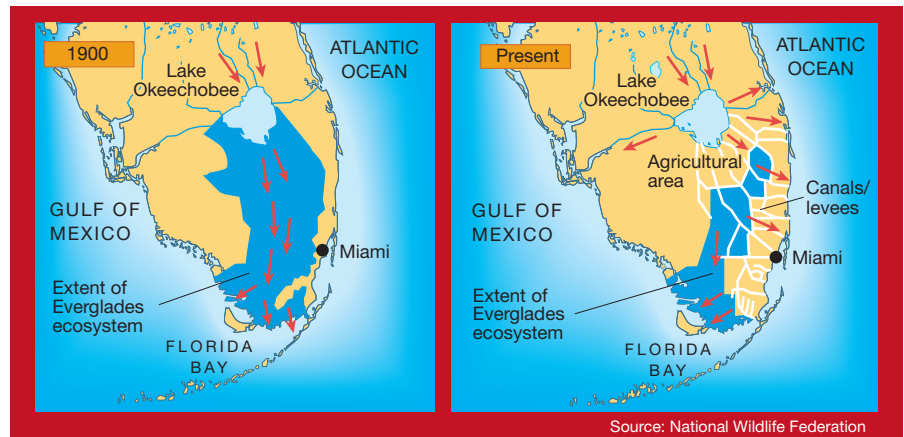
were like before drainage began. Using historical sources such as aerial photos from the 1940s, he has compiled information about past topography and ecology¹.

McVoy's research indicates that the natural flow of water through the Everglades created a series of parallel ridges and sloughs, each about 500 metres wide and with a difference in elevation from ridge top to slough bottom of about half a metre. This pattern provided critical wildlife habitats. The drier ridges supported stands of sawgrass, *Cladium jamaicense*, whereas the sloughs were always flooded, maintaining an aquatic ecosystem. But most of this topography has now gone, and McVoy blames the reduction of free flow.

He and others believe that flowing water kept the sloughs from filling in by slowly transporting organic material downstream in particulate or dissolved forms. To test this hypothesis fully, continuous measurements of such transport at various sites in the Everglades are needed. But preliminary unpublished work by a team led by Daniel Childers at Florida International University in Miami has confirmed that fluffy particulates known as 'flocculent' organic matter are indeed found suspended in the Everglades' waters.

Other researchers have studied soil cores under a microscope, identifying pollen grains to see what type of vegetation dominated particular spots over the past several millennia². They have found evidence for concentrations of sawgrass, characteristic of ridges, and water lilies, characteristic of sloughs, persisting in the same pattern for centuries at a time.

McVoy fears that this original pattern will continue to degrade under the present restoration plan, causing further loss of critical aquatic habitats and possibly allowing dense stands of sawgrass to dominate the



Before and after: about half of the wetland Everglades ecosystem has been lost to drainage.

ecosystem. He has pressed his point at conferences, building a growing ecological consensus behind the view that the restoration plan must be reworked. "He's been standing up and saying if we don't take out these barriers to flow, then we will continue to lose whole landscapes," says Robert Johnson, director of the Everglades National Park's South Florida Natural Resources Center in Homestead. "That's been a hard effort on his part, but I think the scientific community is responding well to what he's saying."

At the Greater Everglades Ecosystem Restoration Science Conference, held in Naples, Florida, last month, ecologists agreed that research to test McVoy's ideas should be a top priority. But it could take 20 years definitively to identify degradation of the natural topography resulting from implementing the plan as it stands—and the critical decisions must be taken now. "The best you can do is look at the pattern and think about the likely processes that formed

it and make likely guesses," says McVoy.

Given the depth of ecologists' concerns, Pimm in January 1999 coordinated a letter to Bruce Babbitt, Secretary of the Interior. Signed by prominent scientists including Edward O. Wilson of Harvard University, the father of biodiversity research, it called for the restoration plan to be reviewed by the National Research Council (NRC), a branch of the National Academy of Sciences. Thanks in part to this letter, such a review is now underway, and the first of a series of NRC reports should be issued by the end of this month.

The architects of the plan defend their work. Tom Teets, project manager for the plan at the SFWMD, and Russell Reed, a senior civil engineer on the project with the Corps of Engineers, say that plans with more barrier removal were modelled, but they led to complications with excessive water or dryness in certain areas.

However, critics argue that options with more barrier removal were not given enough consideration and believe that any problems could be overcome. If the NRC panel backs their views, it will be difficult for the SFWMD and the Corps of Engineers to resist the calls for change. "One thing we always anticipated is constant peer review and interaction with the scientific community as the plan is developed," says Teets.

The rot sets in

Even if the arguments over removing barriers to water flow within the Everglades are resolved, another ecological controversy will remain. This centres on how much adding a large volume of freshwater to the Everglades will affect Florida Bay, into which most of the additional water will drain.

During the 1980s, it became clear that all was not well within the bay, which began to suffer from alarming die-offs of seagrasses—submerged marine flowering plants that stabilize sediments, provide a habitat for fish, crustaceans and molluscs, and are eaten by animals such as turtles and manatees. Blooms of planktonic algae have also become more frequent.



When the levee breaks: engineers are punching holes through some barriers to water flow.

news feature

The most widely accepted theory for why this has happened blames the reduction in freshwater flow from the Everglades caused by their drainage³. The idea's main proponent is Joseph Zieman, a coastal marine ecologist at the University of Virginia in Charlottesville. Historically, he argues, the bay experienced rising and falling inputs of freshwater, and hence salinity, depending on the rainfall in a given year. This variation in salinity supported a diversity of seagrasses.

Once the Everglades began to be drained and the water volume flowing into the bay plummeted, the salinity became more uniformly marine. This allowed a single species, *Thalassia testudinum*, or turtle grass, to take over. During droughts, Zieman argues, salinity now gets high enough in places to damage even the turtle grass. As the dead seagrass rots, releasing nutrients into the water, it feeds algal blooms that in turn block light and cause more seagrass die-offs. The loss of seagrasses also allows storms to churn up more bay sediments, which frees further nutrients to fuel the blooms.

Brian Lapointe of the Harbor Branch Oceanographic Institution in Fort Pierce, Florida, opposes this view. He believes the algal blooms stem in the first instance from elevated levels of nitrogen in water flowing from the Everglades, and that the blooms are responsible for seagrass die-offs.

Lapointe's ideas were long ignored, largely because the prevailing view has been that algal growth in the bay is limited by the amount of phosphorus present, not nitrogen. But in unpublished experiments, Carmelo Tomas of the University of North Carolina at Wilmington and his colleagues have shown that nitrogen can stimulate algal growth in samples taken from parts of the bay. And in a forth-



Out to grass: a shortnose batfish (*Ogcocephalus nasutus*) rests among shoots of turtle grass.

coming paper⁴, Lapointe documents events in the early 1990s when increased freshwater flows into the bay lowered salinity and raised nitrogen levels. This increased algal blooms and seagrass die-offs, even in areas where salinity was not a problem.

Drain storms

Some marine ecologists believe that both Zieman and Lapointe have valid points. "I would say that most people would acknowledge that it's likely that salinity has been an important factor," says David Rudnick, an estuarine ecologist with the SFWMD. "But most scientists look at the whole picture and would say that nutrients have also probably been a factor."

This makes the impact of the Everglades restoration plan on Florida Bay difficult to predict. If salinity is the main problem, then an increase in the volume of freshwater entering the bay should help. But if nitrogen in run-off from the Everglades plays a significant role in the bay's ecological dynamics, then increasing the volume of water flowing into the bay could exacerbate the situation. The present restoration plan aims only to limit phosphorus levels, by controlling inputs from sources such as agricultural fertilizers and incorporating areas of wetland vegetation that will remove the element from the water through natural uptake.

Rudnick and his colleagues have found high levels of some nitrogen-containing compounds in water flowing from the Everglades⁵. But it is not yet clear how much of that nitrogen is natural and how much has arisen from human activity. Until 1998, Don Boesch of the University of Maryland's Center for Environmental Science in Cambridge chaired the Science Oversight Panel for the Florida Bay Program, a consortium of state agencies studying the bay's ecology. He says his group argued that there was a compelling need for more work to investigate nitrogen sources, as well as for systematic tests of the salinity and nitrogen hypotheses. He is frustrated that the agencies working in the bay did not make these studies a higher priority.

Rudnick says that Boesch's message has now been taken on board. "We definitely have more of a focus on nitrogen than we used to have," he says. "I think we will be in a better position to deal with this issue within the next two to three years." But there is little information about nutrient levels and the dynamics of

the seagrass beds during the time before the problems were recognized. Given this, Boesch and Lapointe argue that there is an urgent need for experiments in controlled environments where the effects on seagrass of altering levels of nitrogen and salinity can be studied directly.

If a decision is eventually made to reduce the amount of nitrogen flowing into Florida Bay, it will be more difficult to accomplish than limiting phosphorus. But Lapointe argues that it could be done, for instance by incorporating algal aquaculture into water treatment areas — in essence creating contained algal blooms to soak up the nitrogen. "We should have been researching this for the past decade," he says.

Concerns about nitrogen do not end with Florida Bay. At least some of the run-off from the Everglades makes it through the Florida Keys, the chain of islands that extends from the end of the peninsula, and onto the coral reefs found on the Keys' Atlantic side. Lapointe firmly believes that nitrogen from the Everglades is contributing substantially to the decline of those reefs. Again, Lapointe fears that the Everglades restoration plan could accelerate this decline. "I think he really has something to say," says Phil Dustan, a coral reef ecologist at the College of Charleston in South Carolina, who recently completed a five-year study of Florida's reefs, finding an overall decline in coral cover of nearly 40% over that period⁶.

As the restoration plan unfolds, other problems may emerge. And the performance of the plan's managers and their scientific advisers in wrestling with these issues will be watched keenly by ecologists planning restoration efforts elsewhere. "There's a long road ahead," says Robert Halley, a geochemist with the US Geological Survey in St Petersburg, Florida, who helped develop the restoration plan. "I think we'll stumble along the way, but we'll pick ourselves up."

Mark Schrepe is a science writer in Richmond, Virginia.



Fragile: corals could suffer if levels of nitrogen in run-off from the Everglades increase.

1. South Florida Water Management District *Everglades Consolidated Report 2001* (2001).
2. Willard, D. W., Weimer, L. & Holmes, C. in *Paleoecological Studies of South Florida* (ed. Wardlaw, B. R.) *Bulletins of American Paleontology* No. 362 (in the press).
3. Zieman, J. C., Fourqurean, J. W. & Frankovich, T. A. *Estuaries* **22**, 460–470 (1999).
4. Lapointe, B. E. & Barile, P. *Estuaries* (in the press).
5. Rudnick, D. T., Chen, Z., Childers, D. L., Boyer, J. N. & Fontaine, T. D. *Estuaries* **22**, 398–416 (1999).
6. Dustan, P. in *Seas at the Millennium: An Environmental Evaluation Vol.1* (ed. Sheppard, C. R. C.) 405–414 (Elsevier, Oxford, 2000).

► <http://www.evergladesplan.org>

Sensitive development could protect Amazonia instead of destroying it

Sir— After 40 years of government-induced settlement of the Brazilian Amazon, the core of this region has experienced surprisingly little deforestation. The agriculture and ranching that cause deforestation depend upon reliable roads, which are concentrated along the eastern and southern flanks of Amazonia (Fig. 1). This 'passive protection' of central Amazonia may soon be lost, however, unless the proposed paving of roads through the core of the region is reassessed.

The Brazilian government has revised its directives for Amazonian development as part of its *Avanço Brasil* (Forward Brazil) plan. This is an ambitious programme for development in Brazilian Amazonia, involving infrastructure investments of US\$45 billion over the next eight years.

The plan emphasizes road paving, river channelling, port improvements and expansion of energy production (Fig. 1). If implemented, it will add 6,245 km of paved highways to the region's road network, including the Santarém–Cuiabá and Humaitá–Manaus highways, which cut through the core of the region (Fig 1 A, B)¹. Paving these highways is justified by the transport cost savings that would accrue to soy farmers in north-central Brazil, but would have substantial attendant environmental costs, such as increased deforestation, logging and forest fires. Establishing these 'export corridors' would lead to further dilution of the state's institutional capacity and of the limited resources available for social development within existing Amazon frontiers^{2,3}.

If the historical relationship between roads and forest loss continues, then the planned road paving will cause between 120,000 and 270,000 km² of additional deforestation, and further forest impoverishment through logging and understorey fire over the next two or three decades^{2,3}. Combined with deforestation occurring within the existing frontier, the deforested portion of the Brazilian Amazon could increase from 14% (550,000 km²) today to a third of the total area over the next 20–30 years. This would release 6 billion–11 billion tonnes of carbon into the atmosphere from forest clear-cutting alone³. The roads would come within 50 km of 22 conservation areas and 89 indigenous lands³.

If Amazonia is to avoid the 'business-as-usual' experience of frontier expansion that has consumed most of the world's forests, Brazil should adopt an alternative development pathway in which

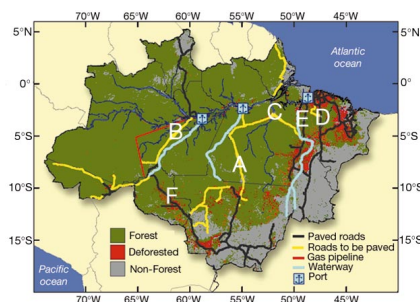


Figure 1 Infrastructure investments described in the *Avanço Brasil* programme. The current frontier is indicated by deforestation (red).

investments are designed to strengthen the economic vigour and institutional capacity of existing settlement regions before roads are paved deeper into the forest.

Investments could be concentrated along the Transamazon, Belém–Brasília, PA-150 and BR-364 highways (Fig. 1 C–F, respectively), where ageing frontiers are languishing in the wake of timber depletion, gold mining and land speculation⁴. These investments include the improvement of local road networks, marketing facilities, technical support, schools and health systems. Financial incentives are needed for sustainable forest management and permanent agriculture initiatives. Institutions must be strengthened to streamline the land titling process, and to implement Brazil's very ambitious environmental legislation through monitoring and control of predatory land-use activities. This approach would stimulate regional development, while preserving the world's largest passive forest reserve.

Georgia Carvalho*†, Ana Cristina Barros†, Paulo Moutinho†, Daniel Nepstad*†

* Woods Hole Research Center, PO Box 296, Woods Hole, Massachusetts 02543-0296, USA

† Instituto de Pesquisa Ambiental da Amazonia, Trav. Eneas Pinheiro 1424, Belém, Pará 66087-430, Brazil

1. Plano Plurianual 2000–2003 (Ministério do Planejamento e Orçamento, Brasília, 1999).
2. *Avanço Brasil: Os Custos Ambientais para a Amazônia* (Instituto de Pesquisa Ambiental da Amazonia, Belém, 2000).
3. Nepstad, D. C. et al. *For. Ecol. Mgmt.* (in the press).
4. Schmink, M. & Wood, C. *Contested Frontiers in Amazonia* (Columbia Univ. Press, New York, 1992).

How electricity could power the car of today

Sir— Although I appreciate Stanford Ovshinsky's kind words for my book, *The Electric Vehicle and the Burden of History* (*Nature* **408**, 289–290; 2000), his review misrepresents an important finding. According to Ovshinsky, my book argues that the historical trajectory of the electric

vehicle was determined by technological constraints — that the "burden of history" means that major technological breakthroughs are now needed to create a market for EVs. In fact, I argue exactly the opposite. The crux of the book is that the future of the electric vehicle (like its past) has relatively little to do with technology *per se* and everything to do with the social context in which owners and drivers choose to use motor vehicles.

The "burden of history" in my book's title refers primarily to the burden of expectation heaped upon the electric vehicle ever since Edison first promised to revolutionize the automobile market with his super-battery in 1898. Such over-promotion has ensured that the electric vehicle is always the car of tomorrow, never the car of today.

The implications of the historical cases explored in the book are clear: policy-makers, manufacturers and drivers must lower their expectations regarding the ability of battery electric vehicles to compete head-to-head with internal combustion cars and look for existing applications where electric vehicles make economic sense.

David A. Kirsch

Anderson Graduate School of Management, UCLA, 110 Westwood Plaza, Los Angeles, California 90095-1481, USA

Did agriculture reduce human lifespan?

Sir— In his review¹ of Clark Spencer Larsen's book *Skeletons in Our Closet: Revealing the Past through Bioarchaeology*, Christopher Wills concludes that "overall health was reduced by ... the introduction of agriculture". He notes that there is little evidence that farmers lived longer than hunter–gatherers.

In the Indian epic *Ramayana*² one finds the following: "In the Golden Age, agriculture was abomination. In the Silver Age, impiety appeared in the form of the agriculture. In the Golden Age, people lived on fruits and roots that were obtained without any labour. For the existence of sin in the form of cultivation, the lifespan of people became shortened." It is conceivable that food shortages in the pre-agricultural era produced healthier individuals because of reduced caloric intake, which is known to delay the onset of age-related pathologies and to extend the lifespan³.

Narendra G. Mehta

3B/33 Takshila, Andheri (east), Mumbai 400 093, India

1. Wills, C. *Nature* **406**, 676–677 (2000).
2. Sen, M. L. (tr.) *The Ramayana of Valmiki* 602 (Munshiram Manoharlal, New Delhi, 1976).
3. Masoro, E. J. *Exp. Gerontol.* **35**, 299–305 (2000).

Tool tests challenge chimpanzees

Do apes fail to understand physics, or do we fail to understand them?

Folk Physics for Apes: The Chimpanzee's Theory of How The World Works

by Daniel J. Povinelli

Oxford University Press: 2000. 407 pp.

£49.50, \$85

Andrew Whiten

As discoveries surfaced that chimpanzees can make tools and show self-recognition along with social politicking and cultural variation, the mental gap between us and them appeared to shrink so far that serious proposals were made to extend human-rights legislation to great apes. In some countries special protection is already in place.

Given this *Zeitgeist*, the results presented by Daniel Povinelli and his colleagues in *Folk Physics* appear quite shocking. They describe 27 meticulously conducted and previously unpublished experiments designed to assess what chimpanzees really understand about the way their physical world works. Seven young chimpanzees were tested on a dozen different kinds of tool use between the ages of five and ten years. Again and again, these juvenile chimpanzees apparently failed to take into account basic aspects of causality, such as that food will not fall 'up' into a container swung from its usual position so as to lie above the food, or that a hook needs to do more than merely touch its target to be a useful tool.

Povinelli's central conclusion is that there is a major, qualitative difference between the everyday 'folk physics' of humans, who can mentally represent unobservable causal factors such as gravity and force, and the chimpanzee's folk physics, which is limited to perceptually available information. Given the significance of this claim, and that Povinelli's group also studies children, it is surprising that we are not reassured by data showing that young children respond differently to the tasks the chimpanzees were set.

Be that as it may, the experiments follow a sustained logic that is fully and clearly explained. If the conclusions are correct, they have far-reaching implications for both chimpanzees and humans. How seriously should we take these results? There is space here to highlight just two main concerns. One is that the juvenile chimpanzee subjects were separated from their mothers in infancy (usually at birth) and reared as a peer group. How would cognitive performance be affected in human subjects reared in the same way? Were inputs that are crucial for a developing chimpanzee absent? Povinelli's answer is that the juveniles' experiences with a variety of tool-based tasks were in all likelihood richer than those of wild chimpanzees. This



Hang on: we may have overestimated chimpanzees' ability to comprehend how things work.

seems a compelling point. Caution remains, however, if primatologist Tetsuro Matsuzawa of Kyoto University is correct in inferring, from his analysis of how skills develop in the wild, that there is a critical period in infancy that is required for the proper, hierarchical development of competent tool use.

A lengthy opening chapter criticizes the 'argument by analogy' — the (mistaken) assumption that if members of two different species behave in similar ways, the same underlying cognitive processes must be at work. I was not persuaded that such criticism offers a radically new insight. Comparative psychologists already know that any behavioural similarity, on further probing, may turn out to be achieved by different cognitive processes. I am equally unconvinced that a fundamental logical flaw exists here, for it would imply that even the deepest probing could never establish both the differences and the similarities in cognition that comparative psychologists seek to delineate.

Perhaps Povinelli has himself fallen victim to the argument by analogy. He illustrates his critique by quoting his earlier studies, which gave a resounding negative verdict to the question of whether chimpanzees understand 'seeing'. This conclusion appears to be based on the analogy that a child failing such tests (understanding the implications of someone covering their eyes) would be unlikely to understand seeing in general. Recently, more natural experiments (involving competition with other chimpanzees over food) were carried out by Brian Hare and his colleagues showing that chimpanzees can and do discriminate important aspects of seeing (and perhaps even the 'knowing' following on from this).

Therefore, we now have to consider that such an analogy does not extend to chimpanzees.

And so to the second major concern. If Povinelli's gigantic prior analysis of chimpanzees' folk psychology can be overturned by an elegant experiment more intuitive for chimpanzees, what of the prospects for the current, equally voluminous onslaught on folk physics? Time will tell. Whatever the answer, this book presents a rigorously documented set of internally consistent results that offer a stalwart challenge for anyone harbouring ambitions to chart the true mentality of chimpanzees. ■

Andrew Whiten is at the School of Psychology, University of St Andrews, Fife KY16 9JU, UK.

Blazing a trail that led to the Moon

Shoemaker by Levy: The Man Who Made an Impact

by David H. Levy

Princeton University Press: 2000. 303 pp.

\$27.95, £15.95

David W. Hughes

The American astrogeologist Eugene Merle Shoemaker (1928–97) was the sort of scientist that most scientists would be proud to be. Shoemaker worked hard all his life, published profusely, clearly enjoyed himself, and left his mark. He was a fountain of ideas, and a man who was only too happy to share these precious thoughts with his many friends and students. ("I'm a professor," Shoemaker told



CORBIS

Deep impact: visiting Meteor Crater as a young student set Shoemaker's career on a new course.

one of his grateful colleagues. "My job is to give away ideas.")

Like many *Nature* readers, he was confronted with the juggling act of modern academia — university duties, research work, grant-proposal writing, data analysis, administration, committee membership all competed for his time. But Shoemaker had a clear view of what he wanted to do, and just did it. Shoemaker was in his early sixties when Levy first met him, but the book under review goes to considerable pains to provide an overview of Shoemaker's early life.

In 1944, aged 16, Shoemaker entered Caltech to study geology. By 1948 he had joined the US Geological Survey and begun his scientific career by specializing in the stratigraphy of the Colorado Plateau and specifically the genesis of its uranium deposits. This was to have been the basis of his Princeton doctoral thesis. But fieldwork in the Rockies had led to a short detour to Meteor Crater on the Arizona–New Mexico border, and the sight of this impact feature was to colour Shoemaker's career from then on.

By 1956, the US government's interest in uranium had moved on to plutonium, and Gene joined a programme called Project MICE. Here, it was hoped, retrievable plutonium would be produced in a Megaton Ice-contained Explosion.

In studying this problem, Shoemaker visited some of the artificial craters that had been produced in the United States by small kiloton nuclear explosions. He noticed that the craters bore an uncanny resemblance to Meteor Crater. Gene returned to Arizona and changed his doctoral topic. His new thesis suggested that Meteor Crater, although nestling in a highly volcanic area, was actually produced by the impact of an iron asteroid. He proved this to be true a year later by his discovery of nearby coesite. This

high-density quartz only crystallizes out under conditions of very high pressure and temperature that are only naturally obtained during the brief moments of impact.

The similarities between impact craters on Earth and on the Moon, and the United States' blossoming lunar exploration ambitions, led to Gene becoming the founder director of the astrogeology programme in the US Geological Survey. NASA then decided not to set up a separate space geology centre; they used Shoemaker's group instead. Gene desperately wanted to fly to the Moon himself and "bang on it with his own hammer". But in 1963 he fell ill, and was eventually diagnosed as having Addison's disease, caused by a malfunction of the adrenal glands. Fortunately, cortisone treatment was just being introduced. Gene was on cortisone pills for the rest of his life, but his dreams of becoming an Apollo astronaut were scuppered.

Nevertheless, Shoemaker almost single-handedly put the science into the US Apollo mission. He was closely associated with the crash-landing Project Ranger, and the soft-landing Project Surveyor, which showed that the lunar soil would support the weight of a spacecraft and an astronaut. But Shoemaker was worried that NASA was concentrating too much on getting men to the Moon and back, and far too little on what they should do while they were there. People were only useful if they could do what a machine could not — setting up and checking out equipment could be done robotically. Astronauts, according to Shoemaker, should be explorers and field geologists. He considered it a waste of time sending humans into space if they were not trained to expect the scientifically unexpected, and to react positively to it.

By 1969, Gene had fallen out with NASA. He left after the Apollo 13 mission and returned to Caltech to chair the geology

department. But he did not enjoy the minutiae of administration, and when he lost interest in a project, he tended to ignore it. He resigned the chair in 1972 and went back to pure research, with the goal of understanding the interrelationship between impacting asteroids and comets and the resulting craters. In 1973, Gene joined Eleanor Helin in setting up the Palomar Planet-Crossing Asteroid Survey, using their 18-inch Schmidt telescope. By 1980, he had been joined by his wife Carolyn, and Eleanor moved to the 48-inch Schmidt.

The Shoemakers were adept at discovering comets and asteroids. Their 'bag' contained 32 of the former and 1,125 of the latter. This harvest was topped by the astounding discovery of P/Shoemaker-Levy 9 on 24 March 1993. The fragments of this comet — pulled apart by tidal forces after approaching Jupiter too closely in July 1992 — hit Jupiter in July 1994.

In June 1984, Gene had returned to his "rock-knocking" fieldwork, spending three months each southern winter in the Australian outback, camping under the stars with his wife, and investigating old impact sites. But on 18 July 1997, he died in a car crash near the border of the Northern Territory and Western Australia.

David Levy joined the Shoemaker Palomar team in July 1989 and quickly became a firm friend and valued colleague. *Shoemaker by Levy* concentrates on this period, and deals in detail with the media interest engendered by the impact of the fragmented Shoemaker-Levy 9 with Jupiter. Levy writes well, and his pacy style keeps his personal story bubbling along superbly. The book is everything a 'good read' should be.

But this will certainly not be the last biography of one of the leading and most colourful of US planetary scientists. And one cannot help feeling that Levy was too close, and that the book has been written too soon. Perhaps by 2050 or 2100 we will see more clearly where Gene Shoemaker fits into the investigation of Armageddon asteroids, killer comets and their potential effects on life on Earth.

It is also too early to argue about whether Gene was the best US planetary scientist of the last century, or number two, three or four on the list. But his life was a fascinating example of how a happy man approaches and successfully dominates a subject he loves. No two scientists are the same, but very few suddenly start working full-time with their wives at the age of 52. And very few, at 52, can throw off the shackles of administration, and turn, full-time, both to searching space for comets and asteroids, and to hunting the desolate ancient regions of the continents for the signatures of the destruction these impactors can inflict on our as-yet defenceless Earth.

Shoemaker by Levy is an interesting biography. But I would love to compare it with,

say, 'Shoemaker by Carolyn', or 'Shoemaker by Eleanor Helin', or even 'Shoemaker by a dispassionate scientific historian in 2050'. ■
David W. Hughes is in the Department of Physics and Astronomy, University of Sheffield, Sheffield S3 7RH, UK.

The psychological toll of battle

A War of Nerves: Soldiers and Psychiatrists 1914–1994

by Ben Shephard

Jonathan Cape: 2000. 480 pp. £20

Hugh Freeman

The limits of human endurance have been the theme of many books, plays, dramas and the visual arts, as well as providing a living laboratory for psychology researchers and scientists. The worst of these situations is warfare, as it involves not just extreme physical stress, but also an agonizing mental conflict between duty and personal survival. Although this has produced a huge literature, there has been no comprehensive overview of the phenomenon of 'shell-shock' and its associated disorders. Ben Shephard has now provided this.

It was not until 1914, with the vast numbers of combatants involved in the First World War, that the fearful effects of battle were seen to become epidemic, and no army had prepared itself for this. Within the first few weeks, soldiers unable to continue fighting were being evacuated to England in large numbers, and 'shell-shock' was the label devised by the men themselves for their condition. "Of all the things that preyed on the nerves and senses," Shephard writes, "shell-fire was the worst." Trench warfare took a terrible psychological toll because it meant "powerless waiting for an impersonal death".

In the medical services, there was no

agreement on either the causes of the problem or what should be done about it. Many of the affected men showed bizarre disturbances — paralysed limbs, loss of speech or sight, or uncontrollable shaking, while others were denied sleep by terrifying nightmares. This 'hysteria' was in fact an unconscious compromise to the dilemma faced by men whose moral code would not allow them to run away but who nevertheless needed to be removed from the trenches. Sigmund Freud had described such a "conversion" of symptoms earlier, but had to admit that his exclusive concern with sexual factors was disproved by the war: fear of death had become the overwhelming influence.

The British Army turned first to neurologists, who had a selective interest in psychiatric disorders and more prestige than asylum doctors. They initially looked for microscopic damage to the brain, but found nothing; the same fruitless search was going on in France and Germany. In all these countries, neurologists used electric-shock treatment to try to dispel the hysterical disturbances. Whatever the rationale for this punitive approach, men generally preferred to return to the front line rather than face the pain of constantly increasing shocks.

A kinder therapy was being developed by British psychologists, notably Charles Myers, together with some psychologically minded doctors. Outstanding among these was W. H. R. Rivers (hero of the first book and the film of Pat Barker's *Regeneration* Trilogy, published by Penguin), whose psychotherapeutic technique — mainly for officers — had a Freudian basis. He accepted that experience not directly accessible to consciousness could have profound effects, and that symptoms were often the expression of mental conflict.

Regimental medical officers were torn between humane concern for their men and the need to keep units up to strength. One officer, trying to gain some respite for a

depleted and exhausted battalion, was told by his army commander that the men showed "an utter want of manly spirit and courage". Yet distinguishing between men suffering genuine terror or numbing from mere cowards or malingerers was a daunting task. A shell-shocked man might be given a wound stripe and a pension, be told to pull himself together, or be shot for cowardice — sometimes it seemed largely a matter of chance which of these it was. Altogether, the British Army executed 307 men, the French 700 and the German Army only 48. After 1918, though, undue sympathy for war neurotics was widely condemned in Germany as contributing to the German defeat. It would be different next time.

But there was a virtual consensus that shell-shocked men who were evacuated far from the battle zone were unlikely ever to return to it. From this experience, the idea of 'forward psychiatry' emerged. Casualties would be dealt with as near to the front line as possible; in many cases, a few days' rest and food were effective enough. Only more serious cases were evacuated further away.

When peace came, every combatant nation had an enormous legacy of psychiatric casualties, with a corresponding burden of pensions — Britain was still paying 40,000 in 1939. Although the war had given a slight boost to the development of psychotherapy in Britain, mental hospitals largely returned to their pre-war torpor.

During the Second World War, psychiatry no longer played second fiddle to neurology — except in the Royal Air Force. Shephard says it is still almost impossible to obtain any information about the degree of psychiatric breakdown that occurred in aircrew. But even rigorous selection among these volunteers failed to pick out every vulnerable one. Meanwhile, the army had to relearn the need for forward psychiatry, finding that when the Allies were winning, psychiatric casualties (and desertions) were far fewer. Psychiatrists also played a major role in the selection of personnel, although without the approval of Winston Churchill. Germany, on the other hand, executed 15,000 men, and over 20,000 more killed or maimed themselves. Whether psychiatric breakdown in war is contagious and whether executions discourage it remain unresolved questions.

The US Army had to start from scratch in learning to deal with psychiatric casualties, which made up some 30% of total casualties when the US Army first encountered the *Wehrmacht* in Tunisia. Vicious fighting in the Pacific also took a heavy toll. The United States had always been generous to veterans, and ample free services were provided for both physical and psychiatric impairment — in Shephard's view, the generosity may often have prolonged the disability.

The Vietnam War was another matter. Morale was low, drug addiction rife (though



Luck of the war: a 'coward' is shot in *Regeneration*; others who acted the same way received pensions.

book reviews

with uncertain effects) and the enemy was everywhere. Many people who fought in the war seemed to develop persisting or even permanent mental damage. The key to understanding this was the nature of the 'stress' — the unresolved threat to the integrity of the organism and its biological coping mechanisms. Vietnam coincided with growing concern in civilian life about the effects of accidents, violence (particularly rape) and abuse of all kinds. The concept of post-traumatic stress disorder was formulated out of all this and is now an inescapable but controversial part of everyday life.

Ben Shephard has exhaustively combed through literature of all kinds concerning psychiatric breakdown in war, although he found both the Falklands and Gulf campaigns deplorably unresearched. His book will be the standard work on the subject for a long time ahead, and it is as remarkable for its humanity and absorbing narrative as for its historical accuracy. It is a notable achievement. ■

Hugh Freeman, former editor of The British Journal of Psychiatry, is at Green College, Oxford OX2 6HG, UK.

Glistering glory and solid economies

The Power of Gold: The History of an Obsession

by Peter L. Bernstein

Wiley: 2000. 432 pp. £17.99, \$27.95

Anthony Vice

The story of gold over the past 5,000 years is a fascinating one, which Peter Bernstein relates with verve and wit. Jehovah selected gold for the tabernacle where humans should come to worship. Croesus bribed the Oracle of Delphi with gold to ensure the security of his rule. The Spaniards took the gold of the New World in a vain attempt to dominate the Old World — many of Ferdinand of Spain's contemporaries would have echoed his command: "Get gold, humanely if possible, but at all hazards — get gold."

Gold has a long and generally successful history as money. The main reasons for this are that it is remarkably durable, is chemically inert and does not easily fragment. It is readily recognized, and a piece of gold is valued only by its weight and purity. Most of the gold that has been mined throughout the world's history still exists — much of it is in the vaults of central banks, some is held by individuals as coin or jewellery, some is in museums where it decorates statues and furniture, and some is at the bottom of the sea in wrecked treasure ships.

The power and appeal of gold are clear, but one wonders whether Bernstein's subtitle is not rather unfair. After all, it served an



Old gold: an employee inspects gold bricks at the Federal Reserve Bank in New York in 1965.

economic purpose in the shape of the gold standard, which lasted two centuries in Britain, not just because of people's obsession, but because linking money to gold brought two great economic advantages. First, the gold standard limited the power of governments to print money and so cause inflation. US President Herbert Hoover put the point with characteristic force to his successor Franklin Roosevelt: "We have gold because we cannot trust Governments." Recall, in this context, that consumer prices have risen roughly 60 times since 1914.

The second great advantage of the gold standard, which showed to such effect between 1870 and 1914, was in bringing certainty to international trade by fixing parities. All the major trading nations were on the gold standard, and so currencies were fixed by their respective gold backing. There was no scope for fluctuations and no possibility of sudden devaluations or revaluations. With tariff barriers coming down, worldwide commerce flourished.

The massive dislocation caused by the First World War strained the gold standard: Britain had lost its supremacy, the German and Russian economies were sidelined and the United States, by then the world's richest nation, decided to bow out of international economics. Many countries, including Britain, nonetheless went back to gold, only to have their plans shattered by the great slump of 1929–31. The adjustments required to remedy the crisis were simply too great for the system. And the gold standard was, at times, brutal. If a currency was overvalued — because of a government deficit used to pay unemployment benefit, for example — gold left that country, bringing down prices and incomes until parity was restored. Governments could check the gold outflow by raising interest rates and cutting expenditure. Many, including Britain, did just that, and brought Europe to the brink of collapse — violent nationalism, large-scale

unemployment and even a mutiny in the British Royal Navy were features of that time.

Yet gold retained one advantage, as was starkly revealed during and after the Second World War. According to the historian Mark Mazower, 1918 saw perhaps a million refugees, but 1945 witnessed ten times that number. The only way these people could get food and clothing was by offering gold, as jewellery or coin. In the central Europe of the 1940s, paper money — even as US dollars or Swiss francs — was simply not acceptable.

Over the past half century, many people have yearned for a world of fixed exchange rates, control over politicians' spending and constant, or even falling, prices. Charles de Gaulle attempted to restore a gold standard in order to outflank the dollar. He called for a doubling of the price of gold, only to find that he had no support and to see his idea buried in the 1968 Paris spring riots.

Bernstein's account of gold's past 20 or so years is particularly good — he was an executive at the Federal Reserve of New York, which handles foreign exchange for the Federal Reserve system. In the late 1970s the gold price roared ahead, reaching an extraordinary \$850 an ounce in 1980. But since then the price has collapsed, and gold has lost any major role in world economics.

President Richard Nixon's New Economic Policy of 1971 set the stage for a revival in gold's role, but the world had come to realize that it could live without gold. The major currencies could fluctuate, but within limits. Politicians also began to realize that electors would vote out governments that let prices rise too sharply. This was a lesson learned quickly in the United States and Germany, but more slowly in the United Kingdom. Perhaps some Britons still hankered after gold — what the nineteenth-century economist Walter Bagehot called "the obvious and natural idol of the Anglo-Saxon". ■

Anthony Vice is at the Institute of Directors, 116 Pall Mall, London SW1Y 5ED, UK.

You can't have it all

Compromise is a major unifying thread in the fabric of contemporary thought.

Neil S. Greenspan

Anyone contemplating the transformation in scientific understanding over the past ten centuries would have to agree that progress has come on numerous and diverse fronts. Is there a conceptual link between these various advances that also distinguishes them from the notions of the past? In subjects as varied as logic, physics, biology and political theory, the acceptance of a fundamental role for limits has grown while the once automatic faith in absolutes has declined. Consequently, an intriguing candidate for a hallmark of modern thinking is the unassuming notion of the trade-off, or compromise.

A logical starting point for considering this thesis is mathematical logic. Kurt Gödel's incompleteness theorem of 1931 established that no finite system of axioms could entail all of the truths that pertain to the arithmetic of natural numbers. To have the deductive power to prove all such truths, one requires additional axioms that inevitably permit the derivation of false statements pertaining to the natural numbers. Therefore, one can have deductive completeness or one can have logical consistency, but one can't have both in an axiomatization of arithmetic.

In physics, trade-offs, although of different sorts, figure in two of the greatest revolutions of the twentieth century. Although special relativity does offer us an absolute — the speed of light in a vacuum — the consequence is the loss of an absolute frame of

reference and the elimination of absolute simultaneity. Shifting focus from the cosmic to the sub-microscopic, Werner Heisenberg achieved fame beyond science by proposing that there is an inevitable trade-off between precision in the knowledge of a particle's position and precision in the knowledge of that particle's momentum. It is from his experience with this quantum realm that Niels Bohr derived his related idea of complementarity. The gist of Bohr's notion is that two perspectives can both be necessary to fully comprehend an entity or phenomenon and yet be mutually incompatible with respect to the experimental means and conditions required for their acquisition.

This notion of complementarity can be useful even at higher levels of organizational complexity. Modern biology is dominated by the effort to relate molecular structure to function. But to determine molecular structure typically requires molecular homogeneity. In contrast, the function of a gene product, or the gene that encodes it, is only derivable from the full range of its interactions with its heterogeneous chemical environment. As Max Delbrück, one of the

In subjects as varied as logic, biology and political theory, the once automatic faith in absolutes has declined.

pioneers of molecular biology, remarked: "It is not to be taken as a forgone conclusion that structure on the molecular scale on the one hand, and integrated function on the other hand are compatible observables."

Given the mechanics of inheritance, and the possibility for environmental change, it should not be surprising to learn that the genomes of organisms are living testimony to compromise. Genes that on average favour reproductive success in one set of conditions often subvert that success under other conditions. The best known human example is the sickle allele at the β -globin locus, which in single dose protects against falciparum malaria but in double dose usually brings life-threatening disease. Another trade-off for which there is evidence is that between human longevity and fecundity. In the wake of the announcement that a first draft of the sequence of the human genome has been completed, there have been grand speculations about engineering immortal humans and the like. Have the speculators considered that some of the genotypic 'improvements' they envision might come with substantial, and potentially cryptic, genetic costs?

Moving from a focus on the individual organism to organized human populations — societies — political thought was for centuries dominated by the dream of the perfect society. Various utopians have assumed that there is ultimately a way to organize society so that everyone prospers, all needs are met, and conflict disappears. Fundamental to this conception is the belief that all societal virtues can be pursued in harmony. In contrast, a more recent strain of thought, as elegantly expounded in the writings of Isaiah Berlin, recognizes that similarly desirable virtues, such as freedom and equality or justice and mercy, can inherently conflict. Success in achieving one virtue can come at the cost of diminishing the other.

The list of trade-offs and compromises inherent in various spheres of human activity could probably be extended indefinitely but, alas, here as well one must balance the desire for completeness against the space limitations of publication. The last word deserves to go to Niels Bohr, who offered what may be the most provocative comment pertaining to the fundamental role of limits, trade-offs and compromises. According to his biographer, Abraham Pais, when Bohr was asked to specify what was complementary to truth he replied: clarity.

Neil S. Greenspan is at the Institute of Pathology, Case Western Reserve University, Euclid Avenue, Cleveland, Ohio, 44106-4943, USA.



Quite contrary: Niels Bohr's coat of arms reflects his view of the trade-offs inherent in quantum theory.

Food, song and speciation

Michael J. Ryan

It is said that you are what you eat. Diet can also determine how you sound and perhaps even what species you are — if you are one of Darwin's finches.

According to the standard view of how new species arise, populations become geographically isolated and adapt to local conditions, and the communication systems used to recognize mates diverge¹. Surprisingly little is known about how mate-recognition systems diverge during speciation². On page 185 of this issue, however, Jeffrey Podos³ describes a clear and direct interaction between ecological adaptation and the divergence of signals that might be used in mate recognition.

Podos's study subjects were Darwin's finches, which inhabit the Galápagos Islands. The diversity of feeding adaptations in these birds, seen in their differing beaks, has been scrutinized by some of the most eminent of evolutionary biologists^{4–6} — first, of course, by Darwin himself. Beaks are adapted to different feeding tasks, from crushing large seeds to using cactus spines as spears (Fig. 1). Yet beak structure not only affects what goes into the bird but also what comes out. Movements of the beak during singing modify both the rate of trills and the range of the frequencies in the song⁷.

Does the size of the beak influence its vocal performance? Podos shows that it does. Species with larger beaks have a more restricted vocal performance, or a greater 'vocal deviation'; that is, they have a narrower frequency range for their trill rate or a slower trill rate for their frequency range. This same pattern was found in an analysis of a single species, the medium ground finch, a bird in which both beak and song structure vary widely.

For many songbirds, the female's preference for the songs of males of the same species, rather than of other species, results in strong pre-mating reproductive isolation⁸. Species stay separate because they don't reproduce with each other successfully. Such reproductive isolation can occur if the two species never mate in the first place, because they don't recognize each other as potential mates; or, if mating does occur, because the partnership results in infertile offspring (such as the mule, the sterile offspring of a horse–donkey mating).

In Galápagos finches it now appears that when beaks become adapted to different food sources, they might simultaneously acquire a vocal signature that distinguishes them from other types of beaks. Females could use this signature in mate choice

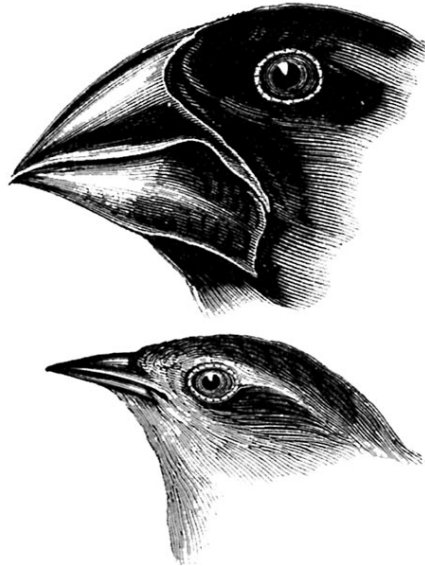


Figure 1 Feeding adaptations. Two species of Darwin's finches, *Geospiza magnirostris* (top) and *Certhidea olivacea*, have very different forms of beak that stem from their different diets. Podos³ argues that song evolution has also been affected, and is a further factor influencing speciation in these birds.

because natural selection should favour their recognition of males with the same beak type — intermediate beaks that result from hybrid matings might be less efficient for dealing with either food type on which their parents specialize. The role of the acoustic structure of trilled song elements in mate choice, however, is not known; this seems to be the next piece of the puzzle which the author needs to address.

The interplay between ecological adaptation, a correlated change in signal structure, and its potential influence on speciation reported by Podos is unusual. Other species, such as cichlid fishes in African lakes, have evolved extremely specialized diets and jaws, together with strong pre-mating isolation that has resulted in an astounding diversity of species⁹. But in cichlids the mate-recognition signal, or at least one of them, is colour pattern¹⁰. So we assume that the more standard pattern of species divergence occurs in circumstances such as these, when signal divergence is apparently not affected by other ecological adaptations.

During the process of speciation, mate-recognition signals can diverge by genetic drift; that is, by chance mutations. But signals

can also adapt to local habitat conditions as other aspects of the organisms' behaviour and morphology adapt to other aspects of the environment. For example, birds on the forest floor produce songs of lower frequency than those in open fields¹¹; chingalos (sparrows) show changes in call structure that match habitat changes with increasing altitude¹²; and a subspecies of cricket frog produces calls that transmit better in its forest environment than do the calls of the other subspecies that inhabits open fields¹³. Similar situations are found in some fishes. The light available for visual displays varies according to the fishes' preferred habitat — say, its favoured depth on a coral reef — and certain species evolve colours and patterns to exploit the prevailing light conditions¹⁴. But the Galápagos finches are unusual in that the ecological adaptation itself is constrained to cause a change in a signal.

Podos suggests that the inextricable link between beak structure, feeding and song has been partly responsible for the Galápagos finches' rapid speciation. But this is an assertion that requires more evidence before it can be accepted. In a larger context, this study further emphasizes how constraints on signals or receivers involved in mate recognition can influence signal or receiver divergence, and thus the probability of speciation. The large number of songbird species might result from the combination of a complex voice box and song learning; the complex voice box allows the production of a greater range of sounds, and song learning promotes cultural evolution of song due to copying errors¹⁵. Variation in the sound-reception organ in the frog's inner ear matches frog species number¹⁶, perhaps because it influences the extent of signal diversity that can evolve^{16,17}.

Podos has revealed an unusual pattern in which ecological adaptation, signal divergence and (perhaps) the opportunity for speciation are linked. It adds a critical piece to our already detailed understanding of natural selection and evolution in this hallmark group of species. But its implications are more far-reaching and general. Podos's study should warn us of the danger of trying to understand adaptation and evolution from a myopic perspective. An organism's phenotype is a complex nexus with different components specialized for different tasks, but in which the degree of optimization of one

component is dependent on others. Perhaps this is why you shouldn't talk with food in your mouth.

Michael J. Ryan is in the Section of Integrative Biology, C0930, University of Texas, Austin, Texas 78712-1064, USA.

e-mail: mryan@mail.utexas.edu

1. Mayr, E. *Animal Species and Evolution* (Harvard Univ. Press, Cambridge, MA, 1963).
2. Rice, W. R. & Hostert, E. E. *Evolution* **47**, 1637–1653 (1993).
3. Podos, J. *Nature* **409**, 185–188 (2001).
4. Darwin, C. *On The Origin of Species* (Murray, London, 1859).
5. Lack, D. *Darwin's Finches* (Cambridge Univ. Press, 1947).
6. Grant, P. *Ecology and Evolution of Darwin's Finches* (Princeton Univ. Press, NJ, 1986).

7. Hoes, W. J. *et al. J. Exp. Biol.* **203**, 1845–1855 (2000).
8. Searcy, W. A. & Marler, P. *Science* **213**, 926–928 (1981).
9. Fryer, G. & Iles, T. D. *The Cichlid Fishes of the Great Lakes of Africa* (TFH, Neptune City, NJ, 1972).
10. Seehausen, O., van Alphen, J. J. M. & Witte, F. *Science* **277**, 1808–1811 (1997).
11. Morton, E. S. *Am. Nat.* **109**, 17–34 (1975).
12. Nottebohm, F. *Condor* **71**, 299–315 (1969).
13. Ryan, M. J., Cocroft, R. B. & Wilczynski, W. *Evolution* **44**, 1869–1872 (1990).
14. Marshall, N. J. in *Animal Signals: Signalling and Signal Design in Animal Communication* (eds Espmark, Y., Amundsen, T. & Rosenqvist, G.) 83–120 (Tapir Academic, Trondheim, 2000).
15. Vermeij, G. *Syst. Zool.* **37**, 69–71 (1988).
16. Ryan, M. J. *Proc. Natl Acad. Sci. USA* **83**, 1379–1382 (1986).
17. Lewis, E. R., Hecht, E. I. & Narins P. M. *J. Comp. Physiol. A* **171**, 421–435 (1992).

molecular cloud allows us a clear view of the surviving cloud core.

There have been many previous attempts to probe the structure and physical conditions of pre-stellar cores of gas and dust by using modern radio telescopes operating at millimetre wavelengths, which are sensitive to emissions from complex molecules^{4,5}. These and other studies have shown that the main ingredient of such cores is molecular hydrogen (H₂) mixed with small but important traces of heavier molecules, along with a sprinkling of dust particles. We also know that temperatures are very low, of the order of 10 degrees above absolute zero. The gas densities of 10⁴–10⁵ cm⁻³ appear high relative to the interstellar environment, but are still at least twenty orders of magnitude smaller than that of their ultimate destination inside stars. Because hydrogen cannot be observed directly at millimetre wavelengths, other trace gases are used to infer gas densities. Unfortunately, this introduces uncertainties in the physical parameters of cloud cores derived from millimetre observations.

In their study, Alves *et al.*¹ have exploited the fact that at infrared wavelengths (a few micrometres) a Bok globule becomes more transparent, even though it is virtually opaque at optical wavelengths. By combining optical and infrared images they measure the degree to which dust particles in the Bok globule B68 block out the starlight of background stars seen through the cloud. They chose to study B68 because it is an almost spherical, isolated globule that is viewed against the rich stellar backdrop near the Galactic centre. So although it is small (0.4 light years in diameter), the light of thousands of stars is attenuated while passing through the dusty globule.

In the same way that the light from the setting Sun becomes redder as it passes through more of the Earth's atmosphere, starlight is reddened as it passes through B68. Alves *et al.*¹ construct a detailed map of how the reddening of light from nearly 4,000 stars changes across the globule, from which they determine the profile of the dust density through the cloud with unprecedented resolution and signal-to-noise ratio. By assuming that the dust is mixed throughout the gas in the cloud, and that the gas-to-dust ratio is 100:1, Alves *et al.* get a good picture of the internal structure of this gas globule. Remarkably, they find a close correspondence between the observed dust profile and theoretical predictions for the structure of a certain type of gas sphere that is stable but close to collapse⁶. With a total mass of only twice that of the Sun, B68 appears to be a prime candidate for a pre-stellar cloudlet.

The formation and evolution of stars represent a slow but inexorable effort by gravity to compress interstellar gas into the increasingly high densities encountered in stellar interiors, and ultimately into the

Astronomy

The dark cradles of stars

Bo Reipurth

Stars are born in dark clouds of molecular gas, which remain shrouded in mystery. Astronomers have found a new way to peer inside these star factories.

One of the success stories of twentieth-century astronomy was the development of a physical understanding of the nature, evolution and death of stars — replacing thousands of years of mystical speculation with firm knowledge. But one piece of the puzzle is still missing: how stars are born. Hidden in the dark interiors of molecular clouds lie the birthplaces of stars, inaccessible at optical wavelengths to the probing eyes of astronomers. Technological breakthroughs in detectors operating at infrared and sub-millimetre wavelengths have improved our picture of stellar genesis in the past two decades, but it is still too early to declare that we understand how stars, including our own Sun, were formed. On page 159 of this issue, João Alves and colleagues¹ bring our knowledge of pre-stellar processes an important step forward.

It is widely agreed that most stars in the Galaxy are born in giant molecular clouds. Hundreds to several thousand stars are produced in these star factories, which convert tenuous gases into objects dense and hot enough to ignite nuclear processes. In the course of a few million years, roughly 10% of a molecular cloud may be turned into stars. But what conditions determine whether a particular clump of gas and dust will become gravitationally unstable and form a star? Such information is critical for calculations of the early stages of star formation.

Alves and co-workers imaged a type of dark cloud known as a Bok globule, at optical and infrared wavelengths, by using telescopes of the European Southern Observatory, including the world's biggest, the Very Large Telescope in the Chilean Atacama desert. Although most stars are born in giant molecular clouds, these enormous stellar nurseries are a mixture of complex environ-



Figure 1 A star-forming Bok globule (GDC1). GDC1 is similar to the B68 globule studied by Alves *et al.*¹, but it has already passed through the collapse phase and formed a Sun-like star in its interior. Although the newborn star is still embedded within thick layers of gas and dust, it reveals its existence by spewing a two-sided supersonic jet into the surroundings. Such jets rid the forming star of rapidly rotating material that would otherwise hinder its formation. Jets also help the newborn star to blow away the remnant cloud material⁸.

ments that are not readily disentangled. Bok globules, on the other hand, are relatively simple, tiny, dense blobs of gas and dust, which can completely blot out the visible light of background stars (see Fig. 1 on page 159). When the eighteenth-century astronomer Sir William Herschel first encountered a Bok globule in his telescope, he exclaimed: "Mein Gott, da ist ein Loch in Himmel" ("My God, there is a hole in the skies")². Bok globules are thought to be the original dense cores of larger molecular clouds, which have been broken up by powerful radiation from nearby massive stars³. The destruction of the rest of the

extreme densities of white dwarfs, and even into neutron stars or black holes. B68 seems to be hovering near the brink of gravitational contraction and will most likely collapse to form a star (Fig. 1), unless stabilized by other forces. Despite its low temperature of 16 K, the primary stabilizing force is thermal pressure. In addition, it is likely that the globule possesses a weak magnetic field, which could help stabilize it further⁷. Nonetheless, B68 appears to be only marginally stable, and so could easily be destabilized by physical processes, such as further cooling or a drop in the internal magnetic pressure, or by being hit by a travelling blast wave from a supernova explosion. So Alves and co-workers

predict that the exquisite dark silhouette of B68 may well, sometime in the future, be converted into yet another little shining star in the Milky Way. ■

Bo Reipurth is at the Center for Astrophysics and Space Astronomy, University of Colorado at Boulder, Boulder, Colorado 80309, USA.

e-mail: reipurth@casa.colorado.edu

1. Alves, J., Lada, C. J. & Lada, E. A. *Nature* **409**, 159–161 (2001).
2. Bok, B. J. *Pub. Astron. Soc. Pacif.* **89**, 597–611 (1977).
3. Reipurth, B. *Astron. Astrophys.* **117**, 183–198 (1983).
4. Benson, P. J. & Myers, P. C. *Astrophys. J. (suppl.)* **71**, 89–108 (1989).
5. Williams, J. P. & Myers, P. C. *Astrophys. J.* **518**, L37–L40 (1999).
6. Chandrasekhar, S. *An Introduction to the Study of Stellar Structure* (Univ. Chicago Press, 1939).
7. Nakano, T. *Pub. Astron. Soc. Jpn* **31**, 697–712 (1979).
8. Heathcote, S. *et al. Astron. J.* **112**, 1141–1168 (1996).

Cancer

Death and methylation

Peter A. Jones

Malignant melanoma cells can resist committing suicide when attacked by chemotherapy. The explanation lies in the discovery that a key gene in the cell-death pathway is switched off in this cancer.

An ability to avoid committing suicide is one of the keys to a cancer cell's survival, whether it is spreading from one part of the body to another by metastasis or facing attack from chemotherapy. Malignant melanoma is a particularly nasty form of cancer in this respect. It is both highly metastatic and resistant to chemotherapy, implying that it has mastered the art of avoiding suicide by not implementing the biochemical pathway that leads to cell death

(apoptosis). How melanomas do this was a mystery until now, because — unlike other cancer cells — melanoma cells usually have fully functional *p53* genes, which are important in triggering apoptosis.

Writing on page 207 of this issue¹, Soengas and colleagues go a long way to solving the riddle. They show that malignant melanoma cells can avoid suicide by inactivating a gene at a step further on from *p53* in the apoptosis pathway. This means that

they can survive during chemotherapy and metastasis even though their *p53* genes are functioning normally. But what makes this study doubly interesting is that the inactivated gene — *Apaf-1* — is merely switched off, instead of being completely lost or mutated. The implication is that the suicide pathway could potentially be reactivated in melanomas, making them less dangerous.

The *p53* protein is a key player in the cell-suicide pathway^{2,3} (Fig. 1, over). When a cell suffers DNA damage or certain types of stress, its *p53* proteins are activated and send a death signal to the cell. *p53* is kept in check by a protein called MDM2, which causes the destruction of *p53*. MDM2 itself is under the control of the *p14^{ARF}* protein, which confines MDM2 to a subsection of the nucleus (the nucleolus). This prevents *p53* breakdown. Stress signals are transmitted to *p14^{ARF}* in part through a death-associated protein kinase, DAPK⁴. This apoptosis pathway is disrupted in most human cancers^{2,3} by downregulation or loss of *p14^{ARF}*, upregulation of MDM2, or mutation (and therefore inactivation) of *p53*.

The exact mechanisms by which *p53* promotes apoptosis are not known, but probably involve Bax, a molecule that stimulates the release of cytochrome *c* from mitochondria. Cytochrome *c* activates the Apaf-1 protein, which in turn activates the enzyme procaspase-9, resulting ultimately in cell death. Several proteins can interrupt this death cascade and thus keep the cell alive. The existence of such a complex pathway (with many feedback loops not shown in Fig. 1) certainly gives plenty of places at

Molecular physiology

Haemoglobin scavenger

Normally, ageing red blood cells approaching the end of their 120-day lifespan are degraded in the bone marrow, liver or spleen. In some circumstances, however — such as during infection with malaria or in some autoimmune disorders — red cells may burst within blood vessels. Haemoglobin (Hb) released from ruptured red cells can be toxic unless cleared rapidly from the circulation. Haptoglobin (Hp), a protein found in blood plasma, binds free Hb, but the events that lead to removal of the Hp–Hb complex from the blood have been elusive — until now.

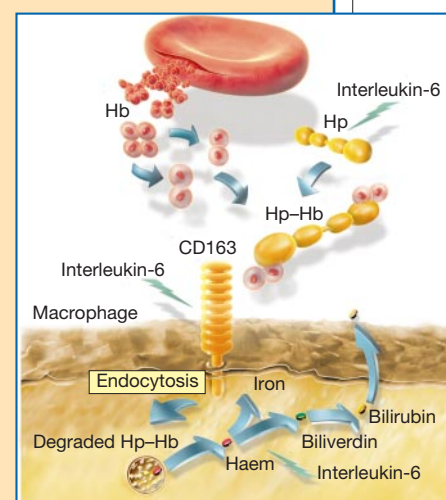
Elsewhere in this issue (*Nature* **409**, 198–201; 2001), Søren K. Moestrup and colleagues describe how they detected a receptor protein that clears the Hp–Hb

complex from blood. They identify the protein as CD163, known to be expressed only on the surface of tissue macrophages and monocytes — white blood cells — and to be involved in inflammation. It seems that the Hp–Hb complex acts as a 'come hither' signal to patrolling macrophages, detected through their CD163 antennae. The Hp–Hb complex is then engulfed by the macrophage and digested to release haem (see graphic).

These findings reveal an intriguing link between iron metabolism and the immune system. The authors speculate that the Hp–Hb complex, like antibodies, may crosslink several CD163 molecules on the surface of macrophages, triggering an internal

signalling cascade that results in increased secretion of anti-inflammatory cytokine molecules. Moestrup and colleagues also propose that different inherited human variants of Hp in complex with Hb may have different anti-inflammatory potency. But it is also feasible that the Hp–Hb complex responds to immune processes. CD163 and Hp are upregulated in response to factors such as interleukin-6, seen in the early, acute phase of inflammation, suggesting that this may be a means of enhancing Hb clearance during inflammatory conditions.

Further evidence is needed. But it is tempting to speculate that variations in Hp and CD163, or perturbations of their function, may be involved in autoimmune



disorders such as systemic lupus erythematosus, and in abnormal iron metabolism. **Carina Dennis**

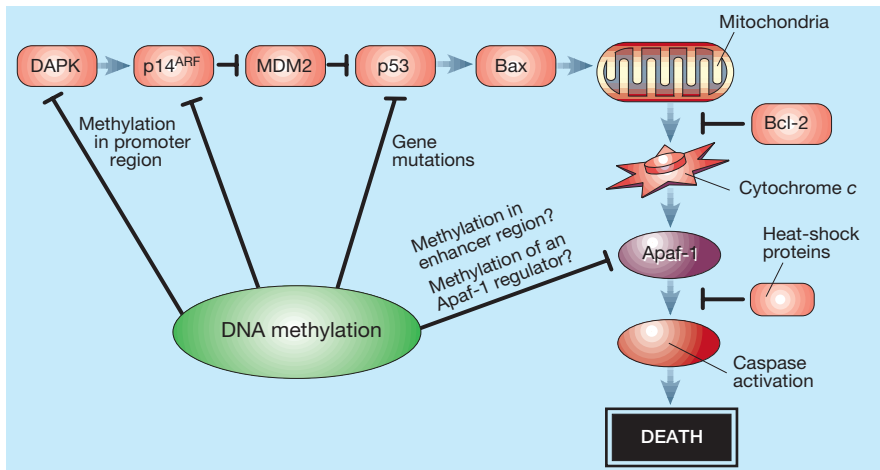


Figure 1 The p53-mediated death pathway, and how it can be subverted by DNA methylation. Increased levels of p53 can lead to the initiation of a cascade of events that results in cell death by apoptosis. p53 is kept in check by a pathway involving DAPK, p14^{ARF} and MDM2. This pathway is disabled in some way in most human cancers, often by inactivation of p53. Malignant melanomas, however, often have intact p53 genes. As shown by Soengas *et al.*¹, these cancers do not express one of the downstream members of the cascade — Apaf-1 — and thus escape apoptosis in response to stresses such as chemotherapy. The Apaf-1 gene seems to be disabled by abnormal methylation. This has also been observed for the DAPK and p14^{ARF} genes, and DNA methylation is involved in generating mutations in the coding region of the p53 gene in other cancers. So DNA methylation can contribute in several ways to inactivating the apoptosis pathway.

which death can be subverted, to which we can now add the inactivation of Apaf-1.

What makes the results of Soengas *et al.*¹ so interesting is that they reveal a reversible switching off — rather than a permanent deactivation — of the Apaf-1 gene. The switch may involve the addition of methyl groups to cytosine nucleotides in DNA, and the removal of acetyl groups from the histone proteins that bundle up DNA into the compressed form seen in the nucleus. Soengas *et al.* show that the Apaf-1 gene can be turned back on by treating cultured melanoma cells with inhibitors of DNA methylation or histone deacetylation.

This 'epigenetic' gene silencing is increasingly being recognized as a common way in which cancer cells inactivate cancer-related genes^{5,6}. Attention has focused on the methylation of cytosines in genes with a high proportion of cytosine-guanosine dinucleotides (the methyl-acceptor sequence) in their promoter regions. After these 'CpG islands' are methylated, and after changes associated with histone deacetylation have occurred, the relevant genes become silent. One example is the abnormal methylation of the promoter of the p14^{ARF} gene. This results in the downregulation of the gene and the degradation of p53, disabling the suicide pathway⁷. The promoter of the DAPK gene is also subject to silencing by methylation⁸.

But unusually Soengas *et al.* find that, although the Apaf-1 promoter is a CpG island (and therefore a potential target for methylation), there were no changes in the methylation of this region before and after melanoma cells were treated with the methylation inhibitor. So the exact mechanism by

which methylation switches off this gene remains in doubt. Perhaps the methylation inhibitor reactivates an unknown gene that controls Apaf-1, or perhaps it demethylates another control region of the Apaf-1 gene, such as an enhancer or insulator.

In any case, it seems that the methylation of cytosine nucleotides in human cancer cells can help to inactivate the apoptosis pathway at several points — either upstream (at p14^{ARF} or DAPK) or downstream (at Apaf-1) of p53. And, as well as contributing to gene silencing, methylation can also substantially increase the occurrence of harmful single-nucleotide mutations when it takes place in the coding regions of genes such as p53 (ref. 9). We clearly pay a price for having methylated cytosine residues in our genomes. Methylation is essential in controlling gene expression under normal circumstances. But it also has several ways in which it can disable the pathways that protect us from cancer.

On the other hand, the fact that the Apaf-1 gene can be silenced by epigenetic changes yet reactivated by drug treatment may have clinical benefits. It might be possible to make melanoma cells sensitive to chemotherapy. Moreover, other important pathways — such as those involving retinoic-acid receptor β 2 (ref. 10), needed for cellular differentiation, or the DNA-repair protein MLH1 (ref. 11) — can be abnormally silenced by DNA methylation. In fact, the list of genes subject to abnormal epigenetic downregulation is growing rapidly. The search for ways to reactivate them will, no doubt, be a major area of research — perhaps involving high-throughput screens such as gene-expression chips — over the next few years. But it may



100 YEARS AGO

In investigating the causes of directions of various spirals, I discovered a certain law and order in the arrangement of the direction of the spiral in horns which will interest many of your readers. (1) That in the antelopes the right-hand spiral is on the left of the head, and the left-hand spiral on the right of the head (crossed). (2) That in sheep the right-hand spiral is on the right of the head, and the left spiral on the left side of head (homonymous, or same name). The wild goats agree with the antelopes in regard to the spiral direction of their horns (crossed), and the oxen agree with the sheep in cases where the spiral can be noted (homonymous). Exceptions are not numerous and not difficult to remember, but this letter is not intended to do more than record the usual rules for spiral directions in horns. If these observations be of value in clearness of description of a difficult point, it will be a gain; and they may also prove useful in classification. By taking a corkscrew (or a right-handed spiral) in the hand, it is easy to verify on the horns themselves the direction of their spiral curves.

From *Nature* 10 January 1901.

50 YEARS AGO

The Société helvétique des Sciences naturelles held its 130th meeting at Davos during August 26–28. The Society has met there twice before—in 1890, when Davos was becoming one of the great health resorts of Europe, and in 1929, when it was also one of the great sport centres... In the afternoon, all sections joined up for the funicular railway ascent of the Weissfluhjoch. The very complex geological panorama was elucidated by Prof. J. Cadisch, of Berne. The Federal Institute for Research on Snow and Avalanches was inspected, with its low-temperature rooms, its thermally tested snowball of about a foot diameter and its stereoscopic microscope showing, in all its perfection, the scintillating branched symmetry of a crystal of snow. Afterwards, a film lecture on "Die Metamorphose des Schneekristalls", by Dr. M. de Quervain, added the dynamic to the static picture, a striking feature being the spontaneous passage to a less symmetrical form with the effect of reducing the surface area. Avalanche research was also illustrated by a film, which by a tragic coincidence had caught a skier being engulfed.

From *Nature* 13 January 1951.

come as a surprise to many scientists that little of our detailed molecular knowledge of how cells function is actually being used at present in treating human cancers.

Peter A. Jones is at the USC/Norris Comprehensive Cancer Center, Keck School of Medicine of the University of Southern California, Los Angeles, California 90089-9181, USA.

e-mail: jones_p@ccnt.hsc.usc.edu

1. Soengas, M. S. *et al. Nature* **409**, 207–211 (2001).

2. Soengas, M. S. *et al. Science* **284**, 156–159 (1999).
3. Vousden, K. H. *Cell* **103**, 691–694 (2000).
4. Raveh, T. *et al. Nature Cell Biol.* **3**, 1–7 (2001).
5. Jones, P. A. & Laird, P. W. *Nature Genet.* **21**, 163–167 (1999).
6. Baylin, S. B. *et al. Adv. Cancer Res.* **72**, 141–196 (1998).
7. Robertson, K. D. & Jones, P. A. *Mol. Cell. Biol.* **18**, 6457–6473 (1998).
8. Kissil, J. L. *et al. Oncogene* **15**, 403–407 (1997).
9. Rideout, W. M. III, Coetzee, G. A., Olumi, A. F. & Jones, P. A. *Science* **249**, 1288–1290 (1990).
10. Widschwendter, M. *et al. J. Natl Cancer Inst.* **92**, 826–832 (2000).
11. Herman, J. G. *et al. Proc. Natl Acad. Sci. USA* **95**, 6870–6875 (1998).

Earth science

In the beginning...

Alex N. Halliday

Grains of the mineral zircon have survived from way back in Earth's history. Analysis of these grains provides information on the state of our planet as long as 4.4 billion years ago.

When did the Earth's continents and oceans first form? On pages 175 and 178 of this issue, Wilde *et al.*¹ and Mojzsis *et al.*² address the question with reports of uranium–lead (U–Pb) ages and oxygen isotopic compositions of extremely old zircon grains. Their results provide evidence that continents and liquid water were surface features of the earliest Earth. Part of one grain appears to have formed 4.4 billion years ago: this is the oldest terrestrial solid yet identified. Although it is unclear how general a picture a few tiny zircon grains can provide, the results represent a significant advance in reconstructing Earth's Dark Ages (Fig. 1).

The U–Pb dating technique depends on the two main isotopes of uranium, ²³⁵U and ²³⁸U, which undergo spontaneous chain decay until they form stable ²⁰⁷Pb and ²⁰⁶Pb, respectively. The probabilities of decay of ²³⁵U and ²³⁸U are extremely small, so their half-lives are long — 700 million and over 4 billion years, respectively. The unique feature that you can simply use the ratio of ²⁰⁷Pb to ²⁰⁶Pb to determine an age, even if the U/Pb ratio has been recently perturbed, makes U–Pb dating the most powerful way of determining the absolute age of material from the early Solar System³. The technique has been used to date the origin of the Solar System at 4.566 ± 0.002 billion years ago⁴.

However, the method is inadequate for defining the rates of formation of planets and their component parts, such as the core and atmosphere. In the 1950s came the first convincing evidence of a now-extinct isotope, ¹²⁹I, that was present at the birth of the Solar System⁵. This discovery sparked the beginning of the use of extinct nuclides, such as ⁵³Mn, ¹⁰⁷Pd, ¹²⁹I and ¹⁸²Hf, which have half-lives that are two orders of magnitude shorter than that of ²³⁵U. Although the parent isotope is extinct, its former abundance can be estimated from a correlation between the

isotopic abundance of the daughter and the parent/daughter element ratio in an object of independently known age. One can then use the isotopic abundance of the daughter in other objects to determine their precise age relative to that of the original object.

From these studies, we know that the accretion (aggregation) of small bodies in the solar nebula occurred within about ten million years of the birth of the Solar System⁶.

Similarly, we can show that Earth's growth was protracted, and dominated by impacts and planetary collisions⁷. By 4.51 billion to 4.45 billion years ago the Earth had reached its present mass, with a metal core and primitive atmosphere^{7–9}. These timescales are consistent with computer models of planetary accretion¹⁰. In its early stages Earth probably had a magma ocean, sustained by heat from impacts and the blanketing effects of a dense atmosphere, but much of that atmosphere would have been lost with the dispersion of the solar nebula and during planetary collisions^{8,9,11}. The Earth would then have cooled quickly, the outer portions would have solidified, and the first primitive crust would have developed.

There was, however, no direct evidence of what such a crust might have looked like. Unlike on the Moon and Mars, no rock older than 4 billion years old seems to have been preserved on Earth. An intense bombardment of the Moon occurred up until roughly 3.9 billion years ago¹². So Earth's earlier crust may have been destroyed by impacts at around the same time, or it could be that a hotter Earth had an inherently unstable surface. Some argued that the earliest crust was like the Lunar Highlands — made from a welded mush of crystals that had previously floated on the magma ocean. Others suggested that it was made of denser rocks like

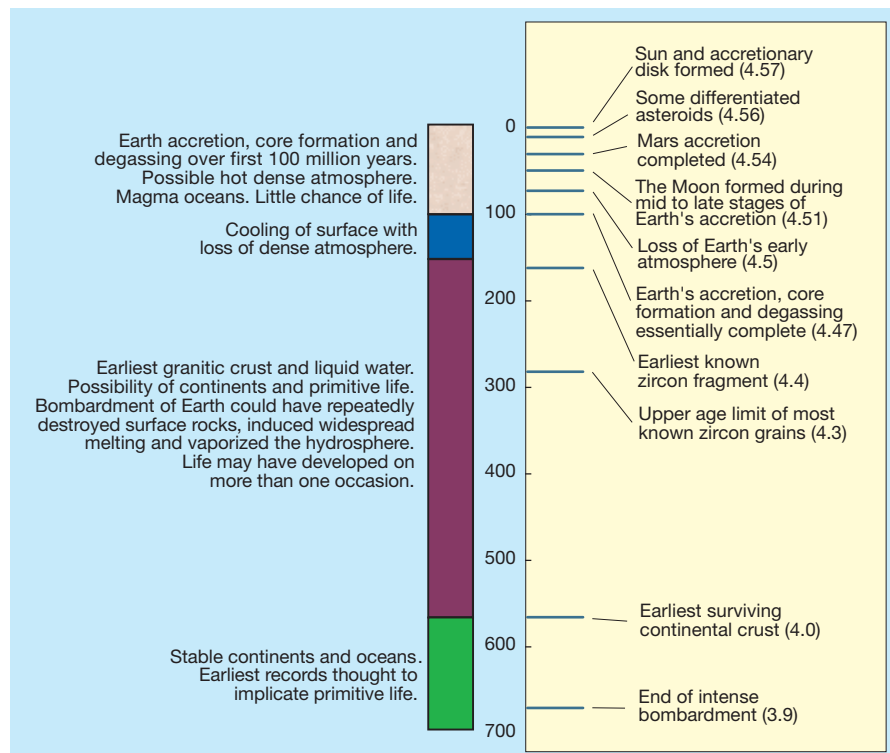


Figure 1 Earth's earliest history, in millions of years, starting from the origin of the Solar System; numbers on the right show absolute ages in billions of years. Dynamic modelling, U–Pb dating of meteorites, and the use of short-lived nuclides all provide evidence for how the Earth formed over the first 100 million years of the Solar System's existence. The Dark Ages of before 4 billion years ago, strictly termed the Hadean, is a period from which no rocks seem to have survived. All we have from this time are a few zircon grains. Wilde *et al.*¹ and Mojzsis *et al.*² are the latest to show that zircons nonetheless represent a wonderful archive of information about the early Earth.

those of the Earth's present ocean floor. But firm evidence was scarce.

In 1983, with the exciting report¹³ of zircon grains over 4 billion years old, we at last had some evidence. But it upheld neither of the above theories. Instead the data constituted evidence that the early crust was like the continental crust today. The original host rocks for the zircons had been destroyed by erosion. But zircon is especially hard and enduring, and the grains survived to tell their tale after becoming incorporated in sedimentary rocks now exposed in Australia. The U–Pb ages of most of the grains turned out to be between 4.1 and 4.2 billion years, with a few grains being slightly older^{13,14}.

But why does the presence of zircons suggest the existence of continental crust? Nearly all zircons grow from granite magmas, which are nothing like those forming the ocean floor or Lunar Highlands. Such magmas usually form at high temperatures and depths of more than 20 km, mainly by melting of pre-existing continental crust above the subduction zones where ocean crust is being dragged down into Earth's mantle. But the granites are buoyant, and rise up to eventually create mountain regions such as the Andes. So the zircons described in the new papers^{1,2} are consistent with the existence of continental crust as far back as 4.4 billion years ago.

We could do with more data to confirm this age, and inferring the existence of entire continents from zircon grains requires quite a leap of the imagination. Nonetheless, large ocean islands such as Iceland contain small amounts of granitic magma, and one can conceive of that magma acting as a nucleus for other material to accrete and create a proto-continent, and then a continent.

Both Wilde *et al.*¹ and Mojzsis *et al.*² also include oxygen-isotope evidence for the presence of liquid water on the early Earth's surface. The oxygen isotopic composition of zircons reflects that of the magma from which they crystallized, which in turn reflects the composition of the rocks that were melted to form the magma. Heavy oxygen (with a high proportion of the isotope ¹⁸O) is produced by low-temperature isotopes between a rock and liquid water. The zircons are rich in heavy oxygen, which implies that the rocks that were melted to form the magma included components that had been at the surface in the presence of liquid water. Earlier on, when Earth was hotter, they might also have formed by melting of once-wet ocean-floor rocks during a process akin to modern subduction. Either way, surface rocks affected by low-temperature fluids were probably being transported to considerable depths and melted. However, there will be suspicions that the ¹⁸O-rich composition was acquired by diffusion after the zircons had been formed, so further information on that possibility is required.

Zircon grains have further potential in

investigations of the early Earth. For example, Wilde *et al.*¹ also analysed trace elements and tiny inclusions to reconstruct the composition of the parent magma, and Hf isotopes can be used to investigate the even earlier history of the Earth¹⁵. And zircons might even contain some isotopic or chemical tracer for early biological activity. For studies such as these we need more grains, and lots of them.

Alex N. Halliday is in the Department of Earth Sciences, ETH Zürich, NO C61, Sonneggstrasse 5, CH-8092 Zürich, Switzerland.

e-mail: halliday@erdw.ethz.ch

1. Wilde, S. A., Valley, J. W., Peck, W. H. & Graham, C. M. *Nature* **409**, 175–178 (2001).
2. Mojzsis, S. J., Harrison, T. M. & Pidgeon, R. T. *Nature* **409**, 178–181 (2001).

Cardiovascular biology

Small cells, big issues

Skip Brass

Identification of a long-sought ADP receptor on platelets explains the effects of two drugs used to prevent strokes and heart attacks. It also points the way to developing other such drugs.

Platelets are the smallest of the human blood cells, measuring about one-tenth the diameter of a white blood cell. They usually circulate freely, maintained in a mobile but inactive state by molecules secreted or expressed by the endothelial cells that line the walls of blood vessels. Beneath the endothelial cells is a protein matrix that includes collagen fibres. When a vessel wall is damaged by injury or disease, platelets adhere to the newly exposed collagen fibres and become active, spreading themselves along the fibres in an effort to close the gap in the layer of endothelial cells (Fig. 1, over). This process — together with the formation of a clot comprising the blood protein fibrinogen — is responsible for stopping bleeding from wounds. At first, platelet activation depends on the presence of collagen and a blood protein called von Willebrand factor. But as platelets accumulate at the site of injury, further activation and the growth of the platelet plug become increasingly dependent on ADP, secreted from platelets or released from damaged red blood cells.

To accomplish all this in a timely fashion, platelets express on their surface a dense array of receptors for molecules, such as ADP, that trigger platelet activation. The molecular identity of two of the ADP receptors is well known. But pharmacological evidence indicated that there must be a third — and, on page 202 of this issue¹, Hollopeter and colleagues identify this hitherto elusive receptor.

Much of the recent work on platelet biology has looked at how platelet activation begins — that is, how non-sticky cells circulating in the bloodstream are transformed into highly sticky ones that adhere to collagen

3. Patterson, C. C. *Geochim. Cosmochim. Acta* **10**, 230–237 (1956).
4. Allègre, C. J., Manhès, G. & Göpel, C. *Geochim. Cosmochim. Acta* **59**, 1445–1456 (1995).
5. Reynolds, J. H. *Phys. Rev. Lett.* **4**, 8 (1960).
6. Lugmair, G. W. & Shukolyukov, A. *Geochim. Cosmochim. Acta* **62**, 2863–2886 (1998).
7. Halliday, A. N. *Earth Planet. Sci. Lett.* **176**, 17–30 (2000).
8. Sasaki, S. & Nakazawa, K. *J. Geophys. Res.* **91**, B9231–B9238 (1986).
9. Porcelli, D., Cassen, P., Woolum, D. & Wasserburg, G. J. *LPI Contrib.* 597 35–36 (Lunar & Planetary Inst., Houston, 1998).
10. Wetherill, G. W. *Annu. Rev. Astron. Astrophys.* **18**, 77–113 (1980).
11. Benz, W. & Cameron, A. G. W. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 61–67 (Oxford Univ. Press, 1990).
12. Wasserburg, G. J., Papanastassiou, D. A., Tera, F. & Huneke, J. C. *Phil. Trans. R. Soc. Lond. A* **285**, 7–22 (1977).
13. Froude, D. O. *et al. Nature* **304**, 616–618 (1983).
14. Compston, W. & Pidgeon, R. T. *Nature* **321**, 766–769 (1986).
15. Amelin, Y. *et al. Nature* **399**, 252–255 (1999).

fibres. This is not a trivial issue. Like many human defence mechanisms, optimal platelet responsiveness represents a balance between too much and too little. A platelet must be activated within seconds of approaching a damaged site, or bleeding will continue. On the other hand, overly responsive platelets could block normal blood flow in the smaller vessels of the heart and brain, particularly at sites where atherosclerosis — the deposition of cholesterol-laden plaques — has narrowed the vessel lumen and damaged the protective layer of endothelial cells. Hence the strong interest in identifying platelet-activating receptors that might serve as targets for drugs designed to prevent the unwanted formation of platelet plugs.

Most of these receptors are coupled to intracellular signalling pathways by molecular switches called G proteins. Human platelets express at least ten G proteins, and each type of receptor can couple to one or more types, initiating signalling through several pathways.

Pharmacological evidence indicates that human platelets express at least three different types of ADP receptor^{2–4}. One type, designated P2Y₁, was thought to be coupled to platelet activation through members of the G_q family of G proteins. This receptor was predicted to be largely responsible for the increase in the concentration of calcium ions inside platelets that is needed to support platelet activation. Indeed, such a role has been confirmed by inactivating the genes encoding P2Y₁ (refs 5, 6) and G_q (ref. 7) in mice. The second receptor — P2X₁ — is a calcium channel, but its contribution to the activation of platelets by ADP has recently

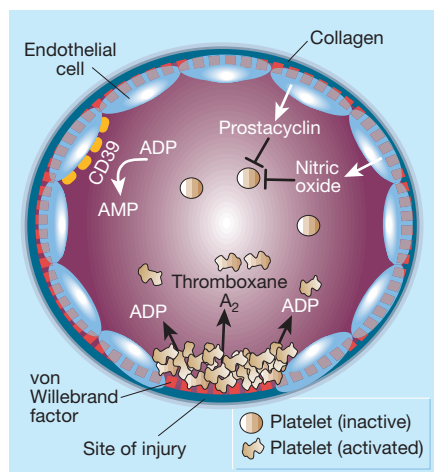


Figure 1 Formation of a platelet plug at sites of blood-vessel injury. Top, circulating platelets are maintained in an inactive state by prostacyclin and nitric oxide, released by endothelial cells. Endothelial cells also express on their surface CD39, an enzyme that converts to inactive AMP any small amounts of ADP that might otherwise activate platelets. Bottom, at sites of vascular injury, endothelial cells are damaged or removed. This exposes collagen fibrils, to which platelets adhere with help from von Willebrand factor, a blood protein synthesized by endothelial cells. Once activated in this way, platelets secrete ADP and thromboxane A_2 . These molecules bind to receptors on circulating platelets, causing them to change shape and become activated, and recruiting them into the growing platelet plug. At the same time, a protein mesh is formed from the plasma protein fibrinogen (not shown). These processes close the gap in the vessel wall, preventing further bleeding until healing can occur.

been questioned⁸. On the basis of circumstantial evidence, the previously unidentified third ADP receptor was predicted to be coupled to the G_i family of G proteins. It is this protein, now dubbed the $P2Y_{12}$ receptor, that has been identified by Hollopeter *et al.*¹.

The G_i family of G proteins is best known for inhibiting the formation of cyclic AMP (cAMP) by the enzyme adenylyl cyclase. But why do this? The answer lies in the balance needed to maintain platelets in an optimal state of responsiveness. Circulating platelets remain inactive in part because endothelial cells secrete prostacyclin (also known as prostaglandin I_2 , PGI_2). This molecule binds and activates receptors on the surface of platelets that stimulate adenylyl cyclase, increasing the formation of cAMP within the platelet. Rising cAMP levels make platelets less responsive to platelet activators. In fact, many such activators — including ADP — work in part by inhibiting adenylyl cyclase and lowering internal levels of cAMP.

The relevance of this effect is shown by the fact that, even before the third ADP receptor was identified, drugs that inhibit the ability of ADP to suppress cAMP forma-

tion in platelets were developed and found to block platelet activation. Two of these drugs, clopidogrel and ticlopidine, reduce the risk of recurrent strokes and heart attacks — catastrophic events that involve platelet activation. In the absence of safe and effective oral drugs that prevent platelets from acquiring the ability to stick to each other, ADP-receptor blockers and aspirin (which inhibits the synthesis of prostacyclin and thromboxane A_2 , another platelet activator) have been used widely to prevent death and disability from heart attacks and strokes.

Unfortunately, some ADP-receptor blockers have occasionally been associated with the development of a life-threatening syndrome characterized by anaemia, kidney failure and, paradoxically, the blocking of small arteries by clumps of platelets⁹. Until now, the development of successors to these drugs has been slow because the adenylyl cyclase-inhibiting receptor for ADP was not known. The cloning of the gene encoding the $P2Y_{12}$ receptor by Hollopeter *et al.*¹ should help considerably, because it defines a target for drug design.

As well as showing that the $P2Y_{12}$ receptor

is present on platelets and can mediate the inhibition of cAMP formation by ADP, Hollopeter *et al.* demonstrate that this protein has the expected pharmacological profile. They also speculate about how a metabolite of clopidogrel might inhibit the receptor, and show that a previously described patient whose platelets fail to respond to ADP lacks a normal form of the gene encoding $P2Y_{12}$. All in all, they make a convincing case that they have indeed identified this biologically and clinically important molecule. ■

Skip Brass is in the Departments of Medicine, Pathology and Pharmacology, and the Center for Experimental Therapeutics, University of Pennsylvania, 421 Curie Boulevard, Philadelphia, Pennsylvania 19004, USA.

e-mail: brass@mail.med.upenn.edu

1. Hollopeter, G. *et al.* *Nature* **409**, 202–207 (2001).
2. Daniel, J. L. *et al.* *J. Biol. Chem.* **273**, 2024–2029 (1998).
3. Jin, J. G. *et al.* *J. Biol. Chem.* **273**, 2030–2034 (1998).
4. Cattaneo, M. & Gachet, C. *Arterioscler. Thromb. Vasc. Biol.* **19**, 2281–2285 (1999).
5. Fabre, J. E. *et al.* *Nature Med.* **5**, 1199–1202 (1999).
6. Léon, C. *et al.* *J. Clin. Invest.* **104**, 1731–1737 (1999).
7. Offermanns, S. *et al.* *Nature* **389**, 183–186 (1997).
8. Mahaut-Smith, M. P. *et al.* *Br. J. Pharmacol.* **131**, 108–114 (2000).
9. Bennett, C. L. *et al.* *New Engl. J. Med.* **342**, 1773–1777 (2000).

Climatology

Glacial hiccups

Didier Paillard

The climate instability of glacial times probably resulted from abrupt switches in ocean circulation. A computer-model simulation provides the first glimpse of the dynamics involved.

Twenty years ago, climate was thought to have remained generally stable, at least for the past few millions of years. The belief was that the ocean and the atmosphere adjusted slowly to stately variations in ice-sheet extent during glacial–interglacial cycles. But the study of palaeoclimatic records has since revealed that climate was in fact highly variable during glacial times¹. It switched abruptly between cold and warm modes, with the temperature in Greenland changing by up to 10 °C in a matter of decades². A crucial step towards understanding these glacial hiccups is presented by Ganopolski and Rahmstorf³ on page 153 of this issue.

It has long been suspected that ocean circulation in the North Atlantic is involved in the abrupt coolings and warmings during glacial periods⁴. The circulation depends mainly on the density of sea water, which is a function of temperature and salinity. These two properties determine the strength of the so-called thermohaline circulation, which in the North Atlantic contributes to the northwards flow of warm water on the surface, followed by heat release and sinking of the cooler water at high latitudes, with ensuing southerly flow at depth.

As early as 1961 it was proposed⁵ that the

interplay between the effects of temperature and salinity could lead to different modes of ocean circulation. In particular, changes in the amount of fresh water in the Nordic Seas can affect deep-water formation in the North Atlantic and alter the thermohaline circulation. As shown in Fig. 1a (overleaf), present-day climate corresponds to an active North Atlantic circulation. If freshwater input into the Nordic Seas rises above a threshold value (F_1 in Fig. 1a), the thermohaline circulation must jump abruptly from its equilibrium (warm) branch to a different one. This new branch corresponds to a much reduced circulation, with colder temperatures at high latitude because less heat is transported there.

When the freshwater perturbation vanishes, the Atlantic circulation does not return to its initial behaviour, but stays inactive. Only a negative perturbation (removal of fresh water, for example by evaporation) can bring it back to normal. In other words, the return pathway is not the same as the perturbation one. The system follows a so-called hysteresis loop as shown in Fig. 1a. Model experiments^{6,7} have confirmed this behaviour for the Holocene — the interglacial of the past 10,000 years. But palaeoclimatic

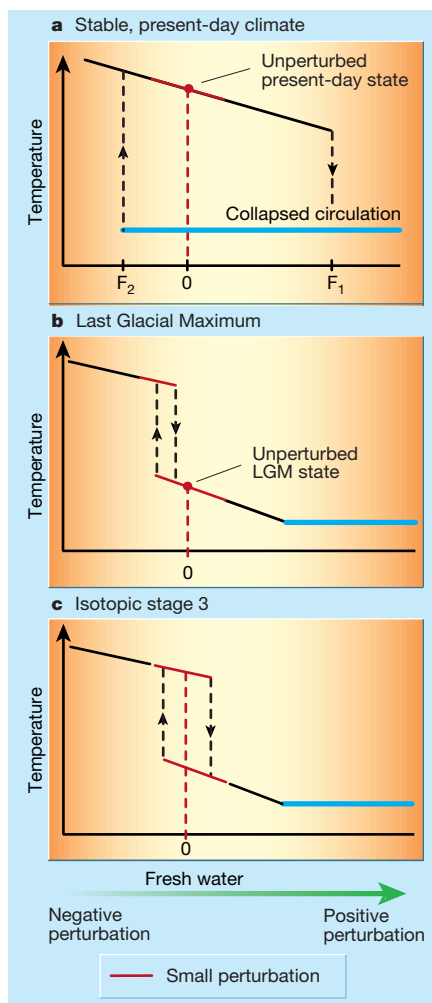


Figure 1 Climate (temperature) stability as a function of freshwater input at high latitudes in the North Atlantic. **a**, Under present-day conditions, North Atlantic climate has essentially two possible equilibria. When freshwater input exceeds a threshold value F_1 , thermohaline circulation jumps from the upper (warm) equilibrium branch to the lower (colder) one, which corresponds to thermohaline collapse (blue line). It can return to the upper branch only if fresh water is removed (by, say, evaporation) and decreases below the threshold value F_2 . The hysteresis width $F_1 - F_2$ is large. So present climate is not destabilized by weak freshwater perturbations. **b**, Under the conditions of the Last Glacial Maximum, the hysteresis is much narrower and so the system is much more sensitive to the input or removal of fresh water — even a slight reduction can induce abrupt warmings, and such Dansgaard–Oeschger warming events are evident in the palaeoclimate record. Large inputs of fresh water, as during Heinrich events (ice-sheet melting), will induce a relatively small cooling through thermohaline collapse. **c**, A guess at an intermediate situation, as pertained during isotopic stage 3, around 50,000–30,000 years ago. The warm (upper) branch is more stable than it is under LGM conditions, corresponding to the longer Dansgaard–Oeschger events that occurred at this time.

records¹ show that the Holocene has been a time of essentially stable climate.

So how do we investigate the large variabilities of glacial times? A prerequisite is a reasonable computer representation of climate at the Last Glacial Maximum (LGM), which occurred around 22,000–19,000 years ago. This is a tough task for the general-circulation coupled ocean–atmosphere models of today's climate; indeed, equilibrating such models for the very different LGM climatic regime is daunting in itself. But a faster track is possible, and a few years ago Ganopolski *et al.*⁸ built an 'Earth model of intermediate complexity' that can perform integrations over thousands of years, yet can also represent the main characteristics of the ocean and atmosphere fairly well. The model's relevance was shown when its results for the LGM climate compared favourably with palaeoclimate reconstructions.

In their new paper³, Ganopolski and Rahmstorf use an improved version of this model to describe the sensitivity of glacial climate to small changes in the amount of fresh water in the Nordic seas. The most notable result is the very different shape of the hysteresis curve under LGM conditions (Fig. 1b). The hysteresis is wide for present-day climate, and can account for the stability of the Holocene because only large freshwater perturbations can destabilize the system. But under LGM conditions, this hysteresis is much narrower, bringing the thresholds for abrupt change closer to the 'unperturbed state'. This explains why glacial climate was so unstable. The LGM equilibrium is located to the right of this hysteresis loop: so even a small loss of fresh water (through increased evaporation, decreased precipitation or run-off from land) induces an abrupt warming, as manifest in the events, known as Dansgaard–Oeschger warmings, recorded in ice cores¹.

The LGM equilibrium is stable if only small amounts of fresh water are added. But, in contrast to the situation today, the lower branch of the LGM hysteresis is not flat. It does not initially correspond to a collapse of the thermohaline circulation, but only to a colder climatic regime where deep-water formation takes place south of Iceland instead of in the Nordic seas. Consequently, there is room for additional cooling through complete thermohaline collapse if there is a substantial addition of fresh water. This is what happened during so-called Heinrich events in the North Atlantic⁹ — times, as reflected in sediment cores, when there was large-scale release, and subsequent melting, of icebergs from the polar ice sheets.

But we also need to simulate these rapid warming and cooling events under different external 'forcings'. To start with, we can make a guess for conditions intermediate between a full glacial and a full interglacial (that is, between Fig. 1a and Fig. 1b). Figure 1c shows

Daedalus

David Jones

David Jones, author of the Daedalus column, is indisposed.

my own guess for such an intermediate period, 'isotopic stage 3' (50,000–30,000 years ago). Here the hysteresis is wider than in LGM conditions and, like today, the unperturbed state lies inside the hysteresis loop. Warm modes would last longer, as is evident in certain Dansgaard–Oeschger events during this time. So the dynamical picture provided by Ganopolski and Rahmstorf³ nicely accounts for the phenomenon of glacial variability. The temporal and geographical patterns of events that emerge from the model, in particular the phase relationship between warmings in Greenland and Antarctica, also compare rather well with the palaeoclimatic data¹⁰.

Plenty of questions remain, of course, most notably that concerning the initial causes of the instabilities. Ganopolski and Rahmstorf carefully avoid the problem by applying a weak, but unexplained, periodic forcing to generate Dansgaard–Oeschger oscillations in the model. They mention solar variability as a possible cause. But little is known of such variability, and invoking it is more pulling out a wild card than providing a solid explanation. The thermal response of the climate model (about 7 °C warming over Greenland) is also weaker than ice-core records indicate (about 10 °C). So other feedbacks probably need to be taken into account — changes in the concentration of atmospheric methane¹⁰, for instance. More broadly, the next step in this line of research will require the coupling of climate models with an ice-sheet model that can simulate the storage and release of large amounts of fresh water over centuries or millennia.

The topic of climate stability is high on both scientific and political agendas, and looks set to stay there. A faithful representation of the Earth system in computer-model form is needed to clarify events of the past, better to predict the future. But the route to that end still lies mostly ahead of us. ■

Didier Paillard is in the Laboratoire des Sciences du Climat et de l'Environnement, Centre d'Etudes de Saclay, Orme des Merisiers, 91191 Gif-sur-Yvette, France.

e-mail: paillard@lsce.saclay.cea.fr

1. Dansgaard, W. *et al.* *Nature* **364**, 218–220 (1993).
2. Severinghaus, J. P. *et al.* *Nature* **391**, 141–146 (1998).
3. Ganopolski, A. & Rahmstorf, S. *Nature* **409**, 153–158 (2001).
4. Broecker, W. S., Peteet, D. M. & Rind, D. *Nature* **315**, 21–26 (1985).
5. Stommel, H. M. *Tellus* **13**, 224–230 (1961).
6. Stocker, T. F. & Wright, D. G. *Nature* **351**, 729–732 (1991).
7. Manabe, S. & Stouffer, R. J. *Nature* **378**, 165–167 (1995).
8. Ganopolski, A. *et al.* *Nature* **391**, 351–356 (1998).
9. Paillard, D. & Cortijo, E. *Paleoceanography* **14**, 716–724 (1999).
10. Blunier, T. *et al.* *Nature* **394**, 739–743 (1998).

An enzymic 'latch' on a global carbon store

A shortage of oxygen locks up carbon in peatlands by restraining a single enzyme.

Historically, northern peatlands have removed carbon dioxide from the atmosphere faster than it has been released, so they now contain 20–30% of the world's soil carbon stock¹ (the equivalent of over 60% of the atmospheric carbon pool²). Here we show that the anaerobic conditions in peatlands prevent the enzyme phenol oxidase from eliminating phenolic compounds that inhibit biodegradation. This indicates that oxygen limitation on a single peatland enzyme may be all that prevents the re-release of a major store of global carbon into the atmosphere, with potentially serious implications for future global warming.

Mechanisms proposed to account for the slow decomposition rates in peatlands include the effects on microbial metabolism of low oxygen availability, low pH, low nutrient supply and low temperatures. But decomposition can be highly efficient in anaerobic environments such as the rumen or man-made sewage digesters. Likewise, it can remain relatively inefficient in fens that are nutrient-rich and whose pH is around neutral, or in mangrove systems that are far from cold³.

One feature common to all of these wetlands, however, is the ubiquity of phenolic compounds. These have attracted intense interest as potent inhibitors of enzymes⁴ and, with the exception of phenol oxidase, few enzymes are able to degrade these recalcitrant materials⁵. But the *in situ* activity of phenol oxidase is severely constrained⁶: it requires bimolecular oxygen, even though it exists in an essentially anaerobic environment. Also, the activity of all the major biodegradative hydrolase enzymes is depressed in peatlands⁷, although they normally retain high activity in anaerobic environments such as the rumen⁸ or anaerobic sludge digest⁹. We propose that the low rate of biodegradation in peatlands is due to oxygen constraints on phenol oxidase, which allow phenolic materials to accumulate and inhibit these pivotal hydrolase enzymes.

We compared enzyme activities under oxygen-saturated and oxygen-free conditions and found that phenol oxidase was the only enzyme to increase in activity under the more aerated conditions (7-fold increase in activity; Table 1). Evidence of a link between phenol oxidase activity and oxygen availability has been found in a Florida wetland, where phenol oxidase activity was detectable only under aerobic conditions⁵. Experimental supplementation with phenol oxidase resulted in a 27% drop in phenolic compound concentrations within 18 hours

Table 1 Effects on enzyme activities

	Control	Manipulated
Effect of oxygen on enzyme activity		
Sulphatase	66 ± 2.3	35 ± 1.4
Phosphatase	571 ± 2.4	387 ± 7.9
β-Glucosidase	237 ± 2.3	177 ± 12
Phenol oxidase	615 ± 93	4,350 ± 27
Effect of increasing phenol oxidase abundance		
Phenolics (μg l ⁻¹)	1,985 ± 55.4	1,444 ± 9.9
β-Glucosidase	1,677 ± 280	10,111 ± 380
Effect of phenolic removal on hydrolase activity		
Sulphatase	579 ± 36	849 ± 43
Phosphatase	3,707 ± 25	4,369 ± 180
β-Glucosidase	1,723 ± 120	2,183 ± 180
Xylosidase	116 ± 2.5	134 ± 5
Chitinase	243 ± 14	296 ± 3.5
Phenol oxidase activity (nmol 2-carboxy-2,3-dihydroindole-5,6-quinone formation min ⁻¹ per g peat), hydrolase activities (nmol methylumbelliferone formation min ⁻¹ per g peat) and phenolic compound concentrations (μg l ⁻¹) are reported as mean ± s.e.		

(Table 1). Lower water-tables, which are associated with increased aeration, also cause a sharp drop in phenolic concentrations¹⁰.

We determined the effect of reduced dissolved phenolic concentrations on hydrolase activities after removing them selectively by using crosslinked *N*-vinyl-2-pyrrolidone¹¹. We found that the activity of hydrolases in peat samples exposed to phenolic-free waters was significantly higher than the corresponding activity in untreated waters containing phenolic compounds at just 2.40 mg l⁻¹ (mean) (Table 1). It has been shown that hydrolases are strongly inhibited (over 80%) in the presence of higher phenolic concentrations¹² or after longer exposure to phenolic materials⁴.

Supplementing aerobic peat samples with additional phenol oxidase led to a significant increase in the activity of hydrolase enzymes (Table 1) in response to the increased removal of inhibitory phenolic compounds. We also showed by doing a regression analysis of data from a 3-month

field survey ($r=0.61$, $P<0.05$) that every doubling in phenol oxidase activity was accompanied by an approximate doubling in CO₂ production.

Taken together, our findings support the idea that oxygen constraints on a single enzyme, phenol oxidase, can minimize the activity of hydrolytic enzymes responsible for peat decomposition. This has profound implications in the context of climate change as a feedback to the process of intensified carbon loss. Increased peat aeration, as a result of droughts predicted by certain climate-change models¹³, has the potential to eliminate a critical mechanism restricting the re-release of CO₂ to the atmosphere. As such, phenol oxidase could be considered to represent a fragile 'latch' mechanism holding in place a vast carbon store of 455 gigatonnes.

Chris Freeman*, Nick Ostle*†, Hojeong Kang*†

*School of Biological Sciences, University of Wales, Bangor LL57 2UW, UK

e-mail: c.freeman@bangor.ac.uk

†Present addresses: Centre for Ecology and Hydrology, Grange-over-Sands LA11 6JU, UK (N.O.); Center for Limnology, University of Wisconsin, Madison, Wisconsin 53706, USA (H.K.)

1. Gorham, E. *Ecol. Appl.* **1**, 182–195 (1991).
2. Oechel, W. C. *et al. Nature* **361**, 520–523 (1993).
3. Painter, T. J. *Carbohydr. Polymers* **15**, 123–142 (1991).
4. Wetzel, R. G. *Hydrobiologia* **229**, 181–198 (1992).
5. McLatchey, G. P. & Reddy, K. R. *J. Envir. Qual.* **27**, 1268–1274 (1998).
6. Pind, A., Freeman, C. & Lock, M. A. *Plant Soil* **159**, 227–231 (1994).
7. Kang, H. J. & Freeman, C. *Soil Biol. Biochem.* **31**, 449–454 (1999).
8. Lee, S. S., Shin, K. J., Kim, W. Y., Ha, J. K. & Han, I. K. *Asian Aust. J. Anim. Sci.* **12**, 988–1001 (1999).
9. Nybroe, O., Jorgensen, P. E. & Henze, M. *Water Res.* **26**, 579–584 (1992).
10. Freeman, C., Lock, M. A. & Reynolds, B. *Chem. Ecol.* **8**, 49–59 (1993).
11. Carpenter, A., Siggla, S. & Carter, S. *Anal. Chem.* **48**, 225–228 (1976).
12. Vuorinen, A. H. & Saharinen, M. H. *Soil Biol. Biochem.* **28**, 1477–1481 (1996).
13. Mitchell, J. F. B. & Warrilow, D. A. *Nature* **330**, 238–240 (1987).

Deafness

Cross-modal plasticity and cochlear implants

Hearing in profoundly deaf people can be helped by inserting an implant into the inner ear to stimulate the cochlear nerve. This also boosts the low metabolic activity of the auditory cortex¹, the region of the brain normally used for hearing. Other sensory modalities, such as sign language², can also activate the auditory cortex, a phenomenon known as cross-modal plas-

ticity. Here we show that when metabolism in the auditory cortex of prelingually deaf children (whose hearing was lost before they learned to talk) has been restored by cross-modal plasticity, the auditory cortex can no longer respond to signals from a cochlear implant installed afterwards. Neural substrates in the auditory cortex might therefore be routed permanently to other cognitive processes in prelingually deaf patients.

Glucose metabolism can be stimulated in the primary auditory cortex and in the auditory-association cortex of prelingually

deaf patients³. In a study of a long-term prelingually deaf patient, the auditory-association cortex was active while watching sign language² but the primary auditory cortex was not. But after this patient received a cochlear implant, the primary auditory cortex was activated by the sound of spoken words, whereas adjacent language areas were not. This phenomenon has been attributed to auditory-to-visual cross-modal cortical plasticity^{2,4}.

We examined and compared glucose metabolism in the auditory and related cortices in 15 prelingually deaf patients before cochlear implantation (mean age, 6 ± 4 years old; age range, 2.2 to 20.3 years); 7 boys and 8 girls) with that of 17 young normal adults (23 ± 3 years old) by using F-18 fluorodeoxyglucose positron emission tomography (PET) and statistical parametric mapping methods. We found that the degree of hypometabolism before the operation was related to the amount of improvement in hearing capability after cochlear implantation (6 subjects received an implant in the right ear, 4 in the left).

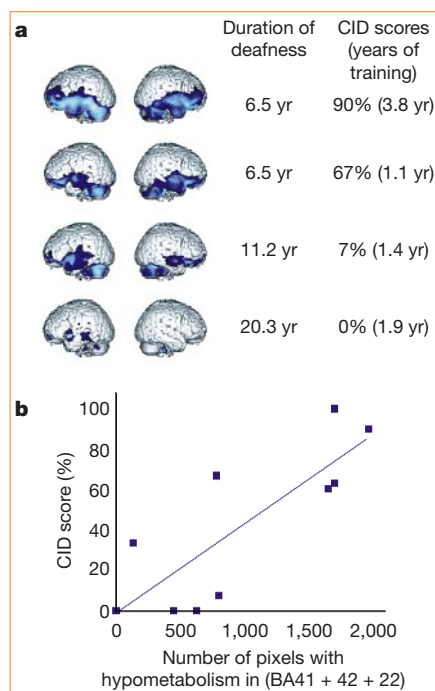


Figure 1 Correlation between hearing response in the prelingual deaf after cochlear implantation and pre-operative hypometabolism in the auditory cortex. **a**, In four representative patients, areas of hypometabolism (shown in blue) on a surface-rendered PET image are displayed with the duration of deafness (age), duration of hearing training, and the CID scores. The 20-year-old patient (bottom row) showed restored metabolism in the auditory and related cortex. She had no hearing capability (CID score, 0%), despite intensive training for 1.9 years. **b**, Scatter plot of CID speech-perception scores and the extent of the hypometabolic area in the primary auditory cortex (BA41) and the auditory-association cortex (BA42, BA22) of both hemispheres in prelingually deaf patients. The pre-operative extent of the hypometabolic areas showed significant correlation with the CID scores after cochlear implantation.

Metabolism fell in the superior temporal (Brodmann area (BA) 41, BA 42, BA 22) or inferior frontal (BA 44 or 45) areas. As the child grew older and the duration of deafness increased, the extent of the hypometabolic area was reduced in the superior temporal and inferior frontal regions (Fig. 1a).

We examined speech-perception performance in a Korean version of the CID (Central Institute for the Deaf) tests during postoperative hearing training, which lasted from 8 months up to 3 years and 8 months, depending on the patient's responsiveness. After cochlear implantation, there was a positive correlation ($r=0.81$, $P<0.005$, $n=10$) between the size of the hypometabolic area in BA 41, 42 and 22 and the hearing-capability score (range was from 0 to 100%; average was 42%) (Fig. 1b). In a multiple regression analysis that included duration of deafness, duration of implant use, and degree of metabolism (measured as the number of hypometabolic pixels) as predictors, the hearing-capability score was predicted independently and most strongly by the degree of metabolism (results not shown).

In our prelingually deaf patients who performed poorly with their cochlear implants, the auditory cortex was probably incapable of perceiving auditory signals from the implants. The realm of afferent neural networks of other sensory systems, such as the visual or somatosensory system, might have increased⁴. Alternatively, higher cognitive functions such as the interpretation of sign language² or lip-reading⁵ could have occupied the relatively under-utilized areas of the auditory cortex. The increased neuronal activity of the auditory cortex, manifesting as an increase in blood flow⁶ and glucose metabolism^{1,3}, implicates cross-modal plasticity as the cause of the continuing lack of hearing response in these prelingually deaf patients.

A similar phenomenon can occur in the congenitally blind⁷, in whom Braille reading activates the visual cortex. When the visual cortex is disturbed by transcranial magnetic stimulation, Braille reading becomes less efficient^{7,8}. Cross-modal plasticity is functionally relevant in congenitally blind patients, but not in patients with acquired blindness⁹.

We have demonstrated the prognostic importance of auditory-to-visual cross-modal plasticity in prelingual deaf patients with the cochlear implants. If cross-modal plasticity restores metabolism in the auditory cortex before implantation, prelingually deaf patients will show no improvement in hearing function, even after successful implantation and concentrated rehabilitation. The resting cortical metabolism of untreated prelingually deaf patients represents a usurping by cross-modal plasticity,

which deters the improvement of hearing capability and the restoration of normal function in the auditory temporal cortices after cochlear implantation.

Dong Soo Lee*, Jae Sung Lee*†, Seung Ha Oh‡, Seok-Ki Kim*, Jeung-Whoon Kim‡, June-Key Chung*, Myung Chul Lee*, Chong Sun Kim‡

Departments of *Nuclear Medicine, †Biomedical Engineering and ‡Otolaryngology, Head and Neck Surgery, Seoul National University College of Medicine, 28 Yungun-Dong, Chongno-Gu, Seoul 110-799, Korea

e-mail: dsl@plaza.snu.ac.kr

1. Naito, Y. *et al.* *Acta Otolaryngol.* **117**, 490–496 (1997).
2. Nishimura, H. *et al.* *Nature* **397**, 116 (1999).
3. Catalán-Ahumada, M. *et al.* *Brain Res.* **623**, 287–292 (1993).
4. Rauschecker, J. P. *Trends Neurosci.* **18**, 36–43 (1995).
5. Calvert, G. A. *et al.* *Science* **276**, 593–596 (1997).
6. Okazawa, H. *et al.* *Brain* **119**, 1297–1306 (1996).
7. Cohen, L. G. *et al.* *Nature* **389**, 180–183 (1997).
8. Sadato, N. *et al.* *Nature* **380**, 526–528 (1996).
9. Cohen, L. G., Ziemann, U. & Chen, R. in *Transcranial Magnetic Stimulation (EEG suppl. 51)* (eds Paulus, W., Hallett, M., Rossini, P. M. & Rothwell, J. C.) 174–182 (Elsevier, New York, 1999).

Molecular engineering

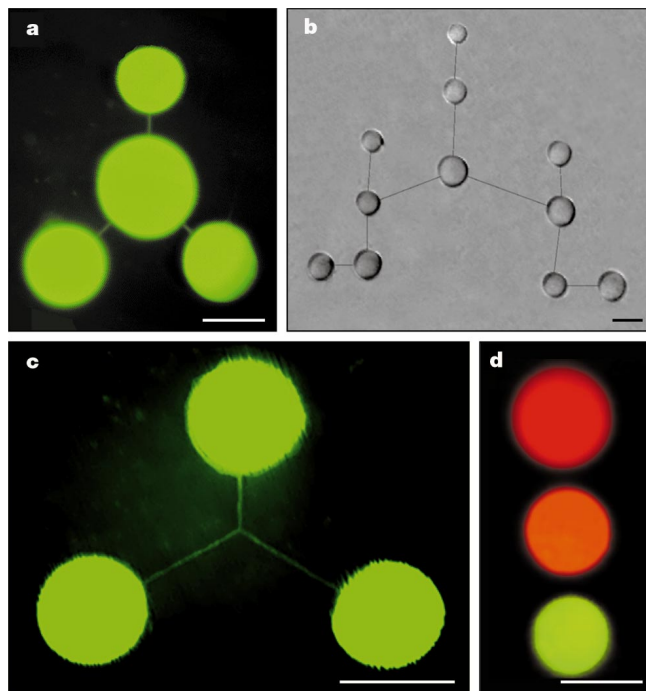
Networks of nanotubes and containers

We have constructed complex two-dimensional microscopic networks of phospholipid bilayer nanotubes and containers in which we are able to control the connectivity, container size, nanotube length, and angle between the nanotube extensions. Containers within these networks can be chemically differentiated and materials successfully routed between two containers connected by a common nanotube. These networks will enable model systems to be devised for studying confined biochemical reactions^{1–3}, intracellular transport phenomena⁴ and chemical computations⁵.

Our method is based on the propensity in liposomes to undergo complex shape transitions as a result of mechanical excitation. It complements earlier micromanipulation protocols for the formation of lipid nanotube networks⁶. As the starting material for generating networks, we used either multilamellar liposomes (5–25 lamellae) prepared by rotary evaporation⁷, or unilamellar liposomes formed by a dehydration/rehydration technique³.

We have been able to produce networks from a wide variety of liposomes of different lipid composition with controlled charge, wettability and functionality. In this instance, we used liposomes either from type XVI-E L- α -phosphatidylcholine isolated from fresh egg yolk or from type II-S acetone-purified soybean lecithin in physiological saline buffer. Liposomes were transferred to borosilicate coverslips coated

Figure 1 Microscopic liposome networks. **a**, Fluorescence image showing a network of four DiO-stained multilamellar liposomes (where DiO is 3,3'-dioctadecyloxycarbocyanine perchlorate) connected by nanotubes. **b**, Bright-field image of an 11-liposome network. As the nanotubes are not visible under bright-field optics, lines have been drawn between liposomes. **c**, Fluorescence image of a symmetric 3-way nanotube junction created by pulling the middle liposome in a V-shaped 3-liposome network away from the other two liposomes. **d**, A 3-liposome network with photochemically differentiated containers created by division of a single Dil-stained liposome (where Dil is 1,1'-dioctadecyl-3,3',3'-tetramethylindocarbocyanine perchlorate) also containing fluorescein. Bottom, liposome depleted of



Dil by a He-Ne laser (633 nm); top, liposome depleted of fluorescein by an Ar⁺ laser (488 nm); middle, unaltered liposome. Micrographs were digitally edited to improve image quality. Scale bars, 10 μm . Further details are available from the authors.

with a thin film of oxidized polystyrene⁸ and examined using an inverted fluorescence microscope.

Nanotubes were formed by mechanical fission of surface-immobilized liposomes (5–20 μm in diameter) by flexible carbon fibres (5 μm diameter, 30 μm long) coated with bovine serum albumin and controlled by high-graduation micromanipulators. A carbon fibre was placed on the equator of a liposome for homofission, resulting in two equally sized nanotube-connected daughter liposomes, or at a 'latitude' for heterofission, resulting in two daughter liposomes of different sizes, and translated in the z -direction (axially) until it touched the surface of the coverslip.

To move two liposomes further apart, we used the micromanipulator-controlled carbon fibre to push on one of the daughter liposomes, typically at a rate of about 5–15 $\mu\text{m s}^{-1}$. The length of tubes could be controlled up to several hundreds of micrometres. We estimated the diameters of unilamellar soybean lecithin and phosphatidylcholine nanotubes to be 200–400 nm from accumulated differential-interference contrast images, although they often appeared to be much narrower, particularly those produced by fission of multilamellar liposomes. It should be possible to customize nanotube diameters by regulating the fission–translation force⁹, the membrane-bending modulus, or the membrane tension⁶.

Using repetitive fission, we could build complex two-dimensional networks by translation of pulled-off vesicles to a target area on the surface. As the surface-adsorp-

tion energies acting on the liposomes are stronger than the nanotube elastic pulling force, daughter liposomes adhere readily and rapidly to the surface as soon as translation is stopped. Tube attachment sites on liposome surfaces are freely diffusible, and nanotubes are always attached at the nearest distance between interconnected daughter liposomes, enabling the angle between two or several tube extensions to be controlled.

The formation of nanotubes from liposome fission is close to 100% successful and takes only a few seconds to complete. Lipid tethers occasionally stick to the electrodes, but these can be removed either by swiftly moving the electrode or by applying a low voltage to the electrode so that the membrane material is reintegrated into the liposome. Networks of four and 11 multilamellar containers, respectively, are shown in Fig. 1a, b. Although multilamellar liposomes can be divided several times, unilamellar liposomes can be split only up to about three times, probably because they have insufficient membrane material to form the nanotube connection.

The complexity of the networks may be increased by creating nanotube junctions such as the symmetric three-way junctions shown in Fig. 1c. We could also make closed or circular networks in which two or several pairs of network-terminating liposomes are electrofused together², enabling large networks of nanotubes and liposomes to be constructed. Networks made from soybean lecithin and phosphatidylcholine liposomes are stable for a couple of hours in physiological saline solution.

It is also possible to manipulate the

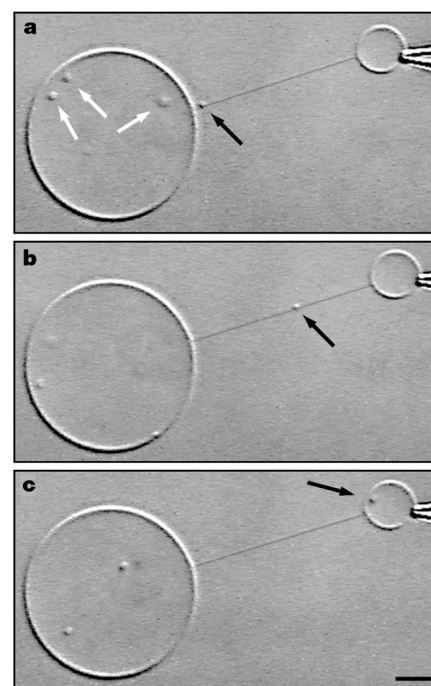


Figure 2 Intratubular particle transport between two unilamellar liposome containers. By microinjecting (continuous flow, 30–50 femtolitres per second) buffer solution into the small liposome, a membrane surface-tension difference between the two containers was created. **a**, A liposome of radius about 150 nm (black arrow) was pulled into the nanotube as a result of the applied tension difference. White arrows indicate several small free-floating multi- and unilamellar liposomes inside the mother liposome. These trapped liposomes form spontaneously in a fraction of liposomes after dehydration/rehydration³. **b,c**, The trapped liposome can be transported all the way to the confines of the connecting liposome. Micrographs were digitally edited and the nanotube contours enhanced to improve image quality. Scale bar, 5 μm .

membrane composition and content of separated liposomes independently after a network has been constructed. By using photochemical, electrochemical, microelectroinjection, electroporation or electrofusion techniques, different enzymes, substrates, dyes, colloidal particles, synthetic and natural organelles, or genetic material, for example, can be introduced into specific containers. We created a network of three liposomes in which the content and membrane composition of each subunit were individually altered by using photobleaching of fluorescent membrane-bound and interior dyes (Fig. 1d). The differentiated liposomes can be sequentially combined into a single container by microelectrofusion (results not shown), a strategy that could be used, say, to initiate chemical reactions in selected containers.

Materials may also be transported between two nanotube-connected unilamellar liposomes by creating a difference in the surface tension of the membrane (Fig. 2), for example by microinjection of buffer solution into one liposome³: this forces lipids to migrate from the liposome with low surface tension to the one with

higher surface tension, which in turn causes a shear-driven movement of the contained fluid. Besides allowing membrane material and small particles to be transported between individual containers, charged materials can also be transported by electrophoresis in addressed nanotubes by application of an electric potential over two nanotube-connected liposomes (results not shown).

In addition to applications relating to chemical and physical processes in cell systems as well as chemical computation, large and complex fluid membrane networks can be used as templates for designing solid-state devices with potential applications in microfluidics and microelectronics^{6,10}.

Anders Karlsson, Roger Karlsson,
Mattias Karlsson, Ann-Sofie Cans,
Anette Strömberg, Frida Ryttsén,
Owe Orwar*

Department of Chemistry, Gothenburg University,
SE-412 96 Göteborg, Sweden

e-mail: orwar@amc.chalmers.se

*Present address: Department of Physical
Chemistry, and Microtechnology Centre, Chalmers
University of Technology, SE-412 96 Göteborg,
Sweden

1. Bucher, P., Fischer, A., Luisi, P. L., Oberholzer, T. & Walde, P. *Langmuir* **14**, 2712–2721 (1998).
2. Chiu, D. T. *et al. Science* **283**, 1892–1895 (1999).
3. Karlsson, M. *et al. Anal. Chem.* **72**, 5857–5862 (2000).
4. Sciaky, N. *et al. J. Cell. Biol.* **139**, 1137–1155 (1997).
5. Laplante, J.-P., Pemberton, M., Hjelmfelt, A. & Ross, J. *J. Phys. Chem.* **99**, 10063–10065 (1995).
6. Evans, E., Bowman, H., Leung, A., Needham, D. & Tirell, D. *Science* **273**, 933–935 (1996).
7. Moscho, A. *et al. Proc. Natl Acad. Sci. USA* **93**, 11443–11447 (1996).
8. Fredriksson, C. *et al. J. Mater. Sci. Mater. Med.* **9**, 785–788 (1998).
9. Waugh, R. E. & Hochmuth, R. M. *Biophys. J.* **52**, 391–400 (1987).
10. Schnur, J. M. *Science* **262**, 1669–1676 (1993).

Marine ecology

Worms start the reef-building process

For reefs to form in wave-swept environments, the sessile organisms that build them need stable foundations^{1,2}. Here we provide evidence that marine worms actively create patches of stable habitat on the sea bed which provide them with food and shelter. In so doing, they bring reef-building organisms together on hard surfaces and so create suitable conditions for reefs to develop.

While maintaining a tropical marine aquarium, we noticed that three small colonies (each weighing 5–20 g) of the scleractinian reef coral *Lobophyllia hemprichii* on the tank's sandy bed had moved 6–16 cm during the night. The corals had become attached by a glue-like substance to a fist-sized lump of reef rock covered in

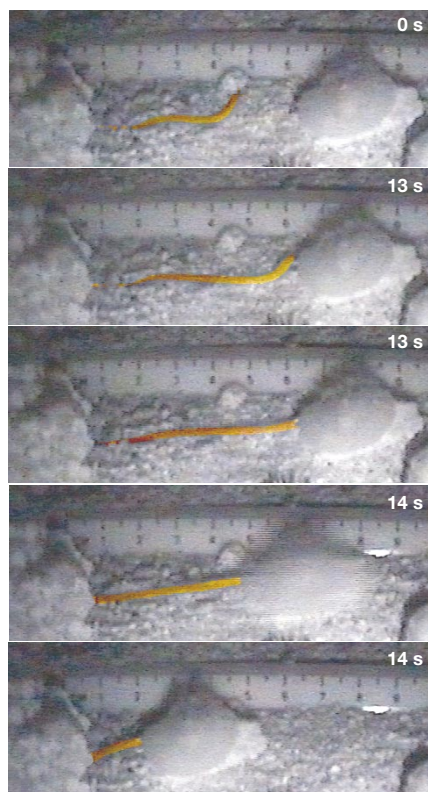


Figure 1 Patch reef construction by a eunicid worm in a reef aquarium. Time-series images, captured from a video recording made under infrared illumination, of the rapid repositioning of a live scleractinian coral from sand bed to solid rock by a eunicid worm for the purposes of habitat assembly. The worm has been coloured to enhance visibility and the images have been adjusted using Adobe Photoshop to improve luminosity and contrast. The ruler in the background is graduated in centimetres and millimetres.

coralline algae. Two colonies were stuck to the side of the rock about 2–4 cm above the sand bed.

We detached the corals from the reef rock and returned them to their former locations. The corals made night-time journeys to the reef rock 21 times over a 1-month period. A 90-degree alteration in the position of the reef rock with respect to the corals, or of the corals with respect to the reef rock, resulted in similar agglomeration around and affixation to the upper or lower sides of the rock.

We made several unsuccessful attempts to film this process under very weak background incandescent white or red light (less than 3 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$) and intermittent weak white or red light (1 minute on, 15 minutes off). Under constant background lighting, there was no movement of corals to the rock. During intermittent light–dark cycles, the corals moved on some occasions, but always during the dark period. This indicated that the corals were being moved rapidly by some photophobic agent.

We next mounted a Hi-8 video camera, equipped with infrared emission, perpendicular to the aquarium overnight. Less

than 1 hour after darkness, the head and upper body of a bootlace-thick eunicid worm appeared from a hole in the rock. The worm sought out a coral and dragged it to the reef rock in less than 2 seconds (Fig. 1).

The persistent nightly efforts of the eunicid worm shown in Fig. 1 to expand and elaborate a durable habitat for itself suggest that such behaviour must have a genetic basis. Although immature, the worm shown in Fig. 1 could raise a 10-g piece of coral above the sand bed of the aquarium and cement it, with its photosynthetic tissue orientated upwards, to the side of a piece of reef rock. Fully grown eunicid worms, which can be more than 2 m long³, must be able to assemble and bind together sizeable mounds of hard substrata.

The fossil record shows that the initial growth of coral reefs depends on the aggregation of sessile, skeleton-forming biota on a hard base^{1,2}. It is not clear how these conditions arise on the shifting seafloors of continental shelf margins. Habitat construction by eunicid and related worms for protection and sustenance is a plausible explanation for both the patchy distribution of benthic communities in general and the development of coral reefs on otherwise unstable substrata¹. By dragging together reef-building organisms and cementing them to mounds of hard substrata, they can assemble spatially heterogeneous, three-dimensional structures that trap sediments¹, enhance settlement and growth of resident species⁴ and other reef biota¹, and provide refuges for cryptic communities^{5,6}.

J. R. M. Chisholm*, R. Kelley†

*Observatoire Océanologique Européen, Centre
Scientifique de Monaco, Avenue Saint-Martin,
MC 98000, Monaco

e-mail: j.chisholm@libertysurf.fr

†Watermark Films, PO Box 1859, Townsville,
Queensland 4810, Australia

1. Wood, R. *Reef Evolution* (Oxford Univ. Press, 1999).
2. Jackson, J. B. C. *Am. Nat.* **111**, 743–747 (1977).
3. Parker, S. P. *Synopsis and Classification of Living Organisms* 2 18–19 (McGraw-Hill, New York, 1982).
4. Jackson, J. B. C. *Bull. Mar. Sci.* **39**, 588–606 (1986).
5. Jackson, J. B. C. & Winston, J. E. *J. Exp. Mar. Biol. Ecol.* **57**, 135–147 (1982).
6. Adey, W. H. & Steneck, R. S. in *The Ecology of Deep and Shallow Coral Reefs* Vol. 2 (ed. Reaka, M. L.) 163–187 (Office of Undersea Research, NOAA, Rockville, 1985).

erratum

Docking of components in a bacterial complex

reply from M. Bochtler, C. Hartmann, H. K. Song,

R. Ramachandran, R. Huber

Nature **408**, 668 (2000)

The meaning of the last sentence was inadvertently altered during editing by introduction of the word “also”, which wrongly gave the impression that our mutational results unambiguously support the EM docking mode. The intention in our reply was to indicate that neither of the two observed docking modes can entirely explain the available mutational and functional data.

Rapid changes of glacial climate simulated in a coupled climate model

Andrey Ganopolski & Stefan Rahmstorf

Potsdam Institute for Climate Impact Research, PO Box 60 12 03, 14412 Potsdam, Germany

Abrupt changes in climate, termed Dansgaard–Oeschger and Heinrich events, have punctuated the last glacial period (~100–10 kyr ago) but not the Holocene (the past 10 kyr). Here we use an intermediate-complexity climate model to investigate the stability of glacial climate, and we find that only one mode of Atlantic Ocean circulation is stable: a cold mode with deep water formation in the Atlantic Ocean south of Iceland. However, a ‘warm’ circulation mode similar to the present-day Atlantic Ocean is only marginally unstable, and temporary transitions to this warm mode can easily be triggered. This leads to abrupt warm events in the model which share many characteristics of the observed Dansgaard–Oeschger events. For a large freshwater input (such as a large release of icebergs), the model’s deep water formation is temporarily switched off, causing no strong cooling in Greenland but warming in Antarctica, as is observed for Heinrich events. Our stability analysis provides an explanation why glacial climate is much more variable than Holocene climate.

Two main types of abrupt climate changes have punctuated the last glacial period: Dansgaard–Oeschger (D/O) events and Heinrich events^{1–8}. D/O events typically start with an abrupt warming of Greenland by 5–10°C over a few decades or less, followed by gradual cooling over several hundred or several thousand years. This cooling phase often ends with an abrupt final reduction of temperature back to cold (‘stadial’) conditions. D/O climate change is centred on the North Atlantic and on regions with strong atmospheric response to changes in that area, and shows only a weak response in the Southern Ocean or Antarctica. The ‘waiting time’ between successive D/O events is most often around 1,500 years, or, with decreasing probability, near 3,000 or 4,500 years (ref. 9). This suggests the existence of an as-yet unexplained 1,500-year cycle which often (but not always) triggers a D/O event.

Heinrich events^{1,5} involve surging of the Laurentide Ice Sheet that covered northern America at the time, and occur in the cold, stadial phase of some D/O cycles. They have a variable spacing of several thousand years. Sediment data suggest that Heinrich events shut down, or at least drastically reduce, the formation of North Atlantic Deep Water (NADW). Records from the South Atlantic Ocean and parts of Antarctica show that the cold Heinrich events in the North Atlantic were associated with unusual warming there (the ‘bipolar see-saw effect’^{10–12}). Sediment data also suggest that changes in the Atlantic thermohaline circulation are crucial in these abrupt climate changes^{13,14}, and it is difficult to imagine a mechanism for such dramatic and rapid temperature changes that does not involve large changes in ocean heat transport.

Many modelling studies in the past have shown that the Atlantic thermohaline circulation (1) is a significant contributor to the regional heat budget over the North Atlantic region, (2) is sensitive to freshwater input into the North Atlantic, and (3) is a nonlinear system with thresholds for transitions between qualitatively different circulation modes. However, previous attempts to simulate the general sensitivity to freshwater input or more specific climate events of the glacial period have been based on perturbations to the modern, rather than to a glacial, modelled climate state. This approach has the obvious limitation that the stability properties of the glacial climate could have been very different from those of the present climate.

Here we present a stability analysis of glacial climate and simulations of D/O and Heinrich events, based on the glacial background climate as simulated in ref. 15. We show that the stability properties of glacial climate differ fundamentally from those of the present climate in our model, and that taking this into

account resolves a number of the paradoxes and questions that emerged from earlier studies. This includes explanations for the characteristic time evolution of D/O events, the different spatial patterns of climate change for D/O and Heinrich events, the recovery of NADW formation after Heinrich events, the fact that cooling pre-dates the iceberg release during Heinrich events and the fact that interglacial climate is much less variable than glacial climate.

Model and background climate

We use the CLIMBER-2 coupled climate model of intermediate complexity^{16,17}, designed for long-term climate simulations. The model has a coarse spatial resolution, allowing us to resolve only the continental-scale features and different oceanic basins. The atmosphere model is based on the statistical-dynamical approach, and does not resolve synoptic variability; the synoptic fluxes are parameterized as diffusion terms with a turbulent diffusivity computed from atmospheric stability and horizontal temperature gradients. The vertical structures of temperature, specific humidity and meridional atmospheric circulation are parametrized. These parametrizations compute the vertical profiles of temperature, humidity and velocity that are used for calculating the three-dimensional advective, diffusive and radiative fluxes. The latter are computed using a multilevel radiation scheme (16 levels) that accounts for water vapour, ozone, CO₂ and the computed cloud cover (stratiform and cumulus). The ocean model is a multi-basin, zonally averaged model similar to that of ref. 18. This model does not resolve the gyre circulation, so that meridional transports of heat and salt are solely due to the meridional overturning circulation and diffusion, except for the North Atlantic, where freshwater transport due to the subpolar gyre circulation is parametrized (see Methods). The sea-ice model predicts ice thickness and concentration and includes ice advection. In the hierarchy of models, our model is placed between energy balance models and general circulation models (GCMs). On a workstation, the model can be integrated for about 10,000 model years in a day.

A number of sensitivity studies (for example, global warming scenarios¹⁹, and reproducing the climate state without NADW formation found in ref. 20) have shown good agreement with GCM results^{15,17}. The model has been used to simulate the climate of the Last Glacial Maximum at 21,000 years before present (21 kyr BP)¹⁵ (for a comparison with other models see ref. 21) and other time slices (Holocene optimum at 6 kyr BP²², Eemian interglacial at 125 kyr BP²³, and glacial inception at 115 kyr BP). A tran-

sient Holocene experiment²⁴ (9 kyr BP to the present) has also been performed.

The glacial climate simulation¹⁵ was forced by prescribed orbital changes, and by changes in atmospheric CO₂ concentrations, continental ice, and sea level. A central result was the change in Atlantic Ocean circulation, most importantly a southward shift in the site of Atlantic deep-water formation. The simulations presented here were performed with an improved version of the CLIMBER-2 model; most notably, the latitudinal and vertical resolution of the ocean model was increased to 2.5° and 20 levels, respectively. Land vegetation cover for the glacial climate was computed by the vegetation submodel^{25,26} of CLIMBER-2. The new model version reproduces all the important features of our earlier glacial simulation.

Stability of the glacial conveyor

Figure 1 shows stability diagrams of the Atlantic circulation with respect to freshwater flux (the most important control parameter) for both modern and glacial climates. Black curves show the standard advective stability diagram as given in ref. 27, where fresh water was added outside the oceanic convection regions; the red curves were computed by adding fresh water directly to the latitudes of the Nordic Seas. The hysteresis loops are narrower in the latter case because a shut-down or start-up of convection is triggered locally before the basin-scale advective stability limits (bifurcation points) are reached.

The stability diagrams reveal a number of important differences between the modern and glacial climates. As in most other models^{20,28}, for present-day forcing (Fig. 1a, c) there are two

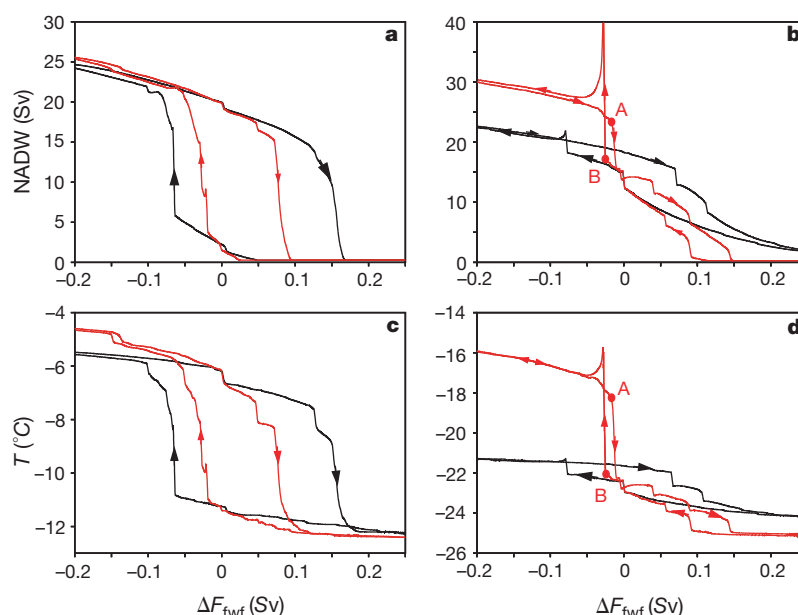


Figure 1 Stability diagrams for the Atlantic thermohaline circulation in the coupled model. The present climate (left panels) differs substantially from the glacial climate (right panels). To calculate the stability diagram the method of ref. 27 was used. The freshwater perturbation ΔF_{fw} was added in the latitude belt 20–50° N to obtain the black curves, and

in the latitude belt 50–70° N to obtain the red curves. Panels **a** and **b** show the ocean circulation response to the freshwater input (in terms of the maximum value of the Atlantic stream function, labelled NADW for North Atlantic Deep Water flow; $1 \text{ Sv} = 10^6 \text{ m}^3 \text{ s}^{-1}$), while panels **c** and **d** show the North Atlantic sector air temperature (60–70° N).

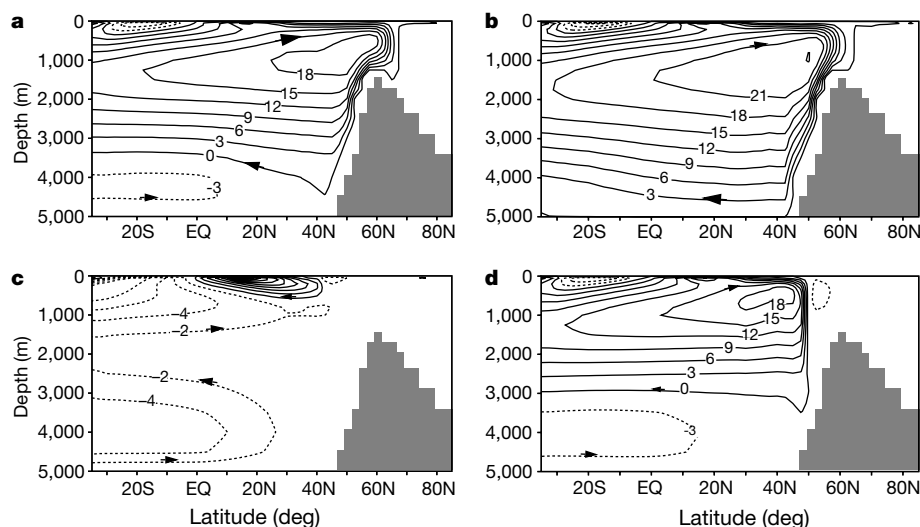


Figure 2 Modes of Atlantic thermohaline circulation in the coupled model. **a**, Holocene 'warm' mode. **b**, Glacial 'warm' (or interstadial) mode. **c**, Holocene 'off' mode. **d**, Glacial 'cold' (or stadial) mode. The diagrams show the stream function (in Sv) for the Atlantic; the

model's bottom topography (representing the sill between Atlantic and Nordic seas) is shaded.

fundamentally different climate modes in the CLIMBER-2 model: with and without NADW formation, that is, the ‘warm’ conveyor belt mode (Fig. 2a) and the ‘off’ mode (Fig. 2c). The ‘width’ of the hysteresis loop in freshwater space is proportional to the oceanic heat transport, as it is the buoyancy gain related to releasing this heat to the atmosphere which has to overcome the buoyancy loss resulting from freshwater input in order to trigger convection. The basin-scale stability threshold, that is, the amount of freshwater input that can be sustained before the circulation breaks down, can be computed from a simple conceptual model²⁸ as

$$F^{\text{crit}} = \frac{\alpha}{4\beta c_p \rho S_0} Q \quad (1)$$

where α and β are the thermal and haline expansion coefficients, c_p and ρ are the heat capacity and density of sea water, S_0 is a reference salinity and Q is the Atlantic heat transport. For a heat transport of 10^{15} W (roughly the present-day heat-transport maximum in subtropical latitudes, both in our model and in observations) this formula results in a hysteresis width of 0.24 Sv, close to the 0.22 Sv seen in Fig. 1a (black curve).

The stability diagrams for the glacial climate have a rather different shape (Fig. 1b, d). Hysteresis ‘width’ is similar because the Atlantic heat transport peaks at 1.0×10^{15} W at 20° N in the glacial mode, just as in the modern mode. But when freshwater inflow to the Atlantic is increased the circulation declines gradually, rather than reaching a clear bifurcation point where the circulation breaks down. When the freshwater inflow is decreased again, the hysteresis behaviour is much less pronounced than in the present climate. The reason is that in the modern climate NADW formation is geographically locked into the Nordic Seas—it cannot jump to the south of the sill near Iceland, because temperatures there are too warm. NADW formation must occur near the sea-ice margin, otherwise it is not dense enough to compete with Antarctic Bottom Water (AABW). In the glacial mode, in contrast, it is cold enough for NADW to form south of Iceland (see Fig. 2d) in the open Atlantic. There it is not geographically locked into place, and the conveyor belt is able to respond to surface freshwater input by gradually retreating southwards and becoming shallower. This leads to a smoother, less nonlinear response compared to the modern conveyor belt. A consequence is that for the unperturbed glacial climate there is no ‘off’ mode and only the ‘cold’ conveyor belt mode (Fig. 2d) is stable.

Another important difference between glacial and modern climates is seen in the response to freshwater forcing in the latitudes of the Nordic Seas (red curves). For a small negative (that is, evaporative) anomaly the circulation jumps to a ‘glacial warm mode’ (Fig. 2b) at point B in Fig. 1. This jump is associated with an increase in Greenland temperature of 5°C (Fig. 1d). The oceanic heat transport across 60° N is 0.12×10^{15} W in this case, and the argument used for the whole Atlantic above can also be applied for the smaller loop of the conveyor belt extending across this latitude into the Nordic Seas in the warm mode: from equation (1) a hysteresis ‘width’ (distance between points A and B) of about 0.01 Sv should apply (taking into account the smaller value of α for the colder glacial waters), as is indeed found in the model experiment. The proximity of A and B to each other in Fig. 1 is thus explained. The reason why the ocean circulation operates only in two distinctly different modes and transitions between these modes can be triggered by small changes in freshwater flux (that is, the reason for the proximity of A and B to zero in Fig. 1) is the glacial surface freshwater flux, which differs substantially from the present one. Owing to a drastic reduction of precipitation and river runoff to the Arctic the freshwater flux is close to zero here, while in the latitude belt 40 – 60° N it is much higher than for present-day conditions. The latter is a consequence of a reduction of evaporation, a southward shift of the storm track and a melting of sea ice in this area. This maximum of the freshwater flux serves as a barrier for the

penetration of Atlantic water masses to the north. This is why convection in the cold (stadial) mode is locked southward of this ‘barrier’ and at the same time a relatively small negative freshwater flux applied to the Nordic Seas can start convection there. As a result NADW formation ‘jumps’ between these locations separated by 1,500 km but cannot be sustained between them.

The different stability properties of the glacial and the modern conveyor belt are thus at least physically plausible. We performed a series of sensitivity experiments to confirm that they are also robust in our model. In these experiments we used a series of different background climates, from full glacial to modern, by varying the prescribed continental ice-sheet size and/or atmospheric CO_2 concentration. We also used a broad range of parameters in the parametrizations of freshwater exchange between the Arctic and the Atlantic (freshwater bypass and sea-ice export, explained in the Methods). These showed that the features shown in Figs 1 and 2 are indeed robust: (1) The glacial climate has a stable ‘cold mode’ (Fig. 2d) with deep water forming south of the Icelandic sill, and a marginally unstable ‘warm mode’ (Fig. 2b) which resembles the modern conveyor belt. (2) The conveyor ‘off’ mode is unstable for glacial conditions.

In addition, the sensitivity study showed that the colder the climate and/or the weaker the export of fresh water from the Nordic Seas, the more points A and B in Fig. 1 move to the left—that is, the more stable the ‘cold mode’ becomes and the larger the freshwater perturbation required to trigger a transition to the ‘warm mode’.

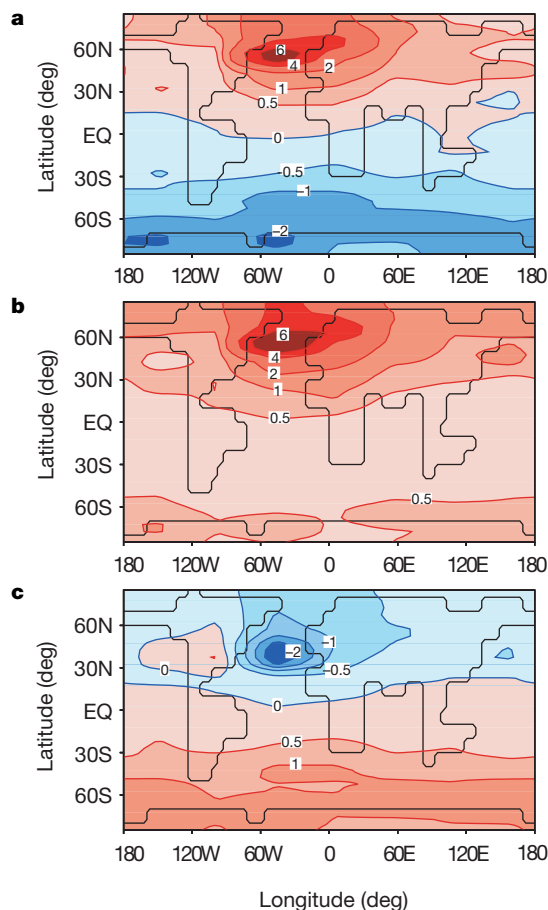


Figure 3 Differences in model-simulated annual mean surface air temperature ($^\circ\text{C}$). **a**, Glacial ‘warm’ mode (Fig. 2b) minus stadial mode (Fig. 2d) in equilibrium. **b**, Warmest phase of a Dansgaard–Oeschger cycle minus stadial phase, 750 years apart (see Fig. 5d). **c**, Conditions during a Heinrich event minus stadial mode (see Fig. 5d). The full time evolution of surface temperature can be viewed as a movie (see Supplementary Information).

The three qualitatively different circulation modes found in the model coincide with the three circulation modes deduced from palaeoclimate data²⁹. The difference in surface temperature between the warm and cold glacial modes is shown in Fig. 3a; it is centred on the northern North Atlantic where it exceeds 6 °C. The Atlantic oceanic heat transport in the glacial warm mode peaks at 1.3×10^{15} W at 20° N compared to 1.0×10^{15} W for the cold mode (Holocene warm mode: 1.0×10^{15} W), but more important for the warming is the fact that the heat transport reaches further north in the warm mode and reduces the area of sea ice in the Northern Hemisphere by one-third.

Dansgaard–Oeschger events

Spectral time-series analysis^{30,31} of Greenland ice-core data (Fig. 4) reveals a dominant peak at a periodicity of 1,470 years associated with the D/O events. Our hypothesis is that a weak cycle (of unknown origin) with this period exists in the climate system, and that this cycle could trigger transitions in the Atlantic Ocean circulation. To test this hypothesis, we imposed a periodic variation in the freshwater forcing of the Atlantic in the latitude belt 50–80° N (Fig. 5a). The amplitude of this forcing is very small (about 30 cm yr⁻¹ in surface flux or 0.03 Sv in total, compare to Fig. 1); it could, for example, represent changes in river runoff, sea-ice export, shifts in the Atlantic storm track or changes in mass balance of the adjacent ice sheets.

The response to the imposed forcing is shown in Fig. 5. After a period of decreasing freshwater input into the region corresponding to the Nordic Seas, convection is triggered there and a sudden incursion of warm, salty Atlantic water occurs. This represents a flip from the ‘cold’ to the ‘warm’ conveyor belt mode (transition B in

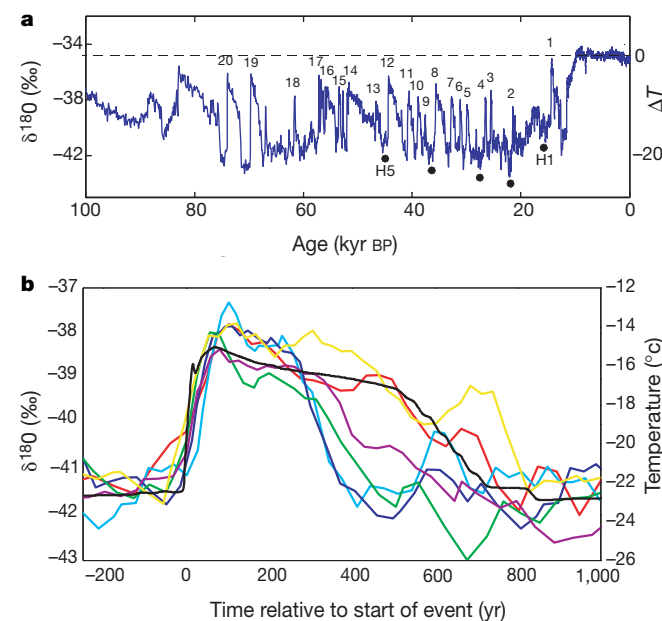


Figure 4 Abrupt climate changes in Greenland ice-core data. **a**, Record of $\delta^{18}\text{O}$ from the GRIP core⁴⁶, a proxy for atmospheric temperature over Greenland (approximate relative temperature range⁴⁷ (in °C) is given on the right). The glacial climate is punctuated by Dansgaard–Oeschger (D/O) warm events (numbered). The timing of Heinrich events H1–H5 is marked by black dots. **b**, Time evolution of recent D/O events taken from **a** (no. 3, light blue; no. 4, dark blue; no. 5, purple; no. 6, green; no. 7, orange; no. 10, red). Many D/O events show the characteristic slow cooling phase after the initial warming, followed by a more abrupt temperature drop. Some events are much longer but still show this general characteristic (for example, nos 8, 12, 19, 20). For comparison, a modelled D/O event is shown in black. For the model we show the North Atlantic sector air temperature 60–70° N (scale on the right), which is a proxy for Greenland temperature in our coarse-resolution model.

Fig. 1b). As usual when convection is triggered after a period of stagnation, stored potential energy is released from the water column leading initially to a vigorous flush. Given that the ‘warm’ mode is not stable under glacial conditions, it gradually decays over the next several hundred years. Finally, a threshold is crossed where convection in the latitudes of the Nordic Seas stops and the system falls back into the stable ‘cold’ mode. It remains there until the next event is triggered.

This cycle has a characteristic salinity signal (Fig. 5c) in the high latitudes as a result of the salt-water incursions, which far exceeds the direct salinity changes due to the imposed freshwater forcing and which matches the observed salinity variations at this latitude³².

A characteristic time evolution of Greenland temperatures is associated with the simulated sequence of events (Figs 4b, 5d): an abrupt initial warming, then a gradual cooling trend terminated by a rapid temperature drop back to stadial conditions. Many of the observed D/O events have similar characteristics (Fig. 4b). In Antarctica, temperatures are increasing during the stadial phase and decreasing during the warm event, but the amplitude of the response is small.

It is important that the characteristics of the simulated D/O events do not depend on the imposed forcing cycle. The forcing only acts as a trigger; once an event is set off it follows its own internal dynamics. If the same experiment is performed with different amplitudes of the forcing cycle, then the threshold behaviour becomes clear: for an amplitude of 0.015 Sv no D/O events are triggered and the Greenland temperature remains constant, whereas for a forcing amplitude of 0.045 Sv the events evolve in the same way and with the same amplitude as shown for 0.03 Sv (Fig. 5b–e), except that they are triggered slightly earlier in the cycle.

The surface temperature response (defined as the temperature of the warm phase of a cycle minus the temperature of the cold phase of that cycle) is shown in Fig. 3b. Like the equilibrium difference between the ‘warm’ and ‘cold’ conveyor belt modes (Fig. 3a), the region of maximum response is centred on the northern North Atlantic. However, owing to the thermal inertia of the oceans the transient response has the same sign globally, and there is stronger D/O warming in the subtropical Atlantic, in better agreement with palaeoclimatic data from this region. In the Southern Hemisphere the effect of the D/O cycles is weak.

Heinrich events

To simulate Heinrich events a much larger freshwater perturbation was added to the northern North Atlantic (amplitude up to 0.15 Sv, Fig. 5a), which mimics the effect of a major ice-sheet surge³³. In response to this freshwater release, the conveyor belt shuts down and the system makes a transition to the ‘off’ mode of Atlantic circulation (Figs 5b, 2c). As the stability diagram (Fig. 1b) shows, this is not a stable circulation mode under glacial conditions, so that the conveyor belt restarts spontaneously after the freshwater influx comes to an end.

What is most interesting is the surface temperature response in Greenland and in Antarctica (Fig. 5d, e). At the start of the Heinrich event the circulation is already in the stadial mode, and Greenland is already cold and beyond the reach of the conveyor belt, so that in Greenland there is hardly any further cooling. This is an important agreement with the Greenland palaeo record (Fig. 4), which shows similar stadial temperatures irrespective of the occurrence of Heinrich events. The strongest cooling caused by the conveyor-belt collapse occurs further south in the subtropical Atlantic³⁴ (Fig. 3c). This is supported by palaeoceanographic sea surface temperature data from the Mediterranean³⁵ and the Atlantic off Portugal³⁶, where Heinrich events register much more strongly than D/O events. The same is true for Antarctica: Heinrich events show the bipolar see-saw with a much stronger Antarctic response, which is due to the dramatic drop in interhemispheric heat transport resulting from the collapse of the conveyor belt³⁷. The sequence of

events in Antarctica shows a gradual warming during Greenland stadials and a cooling during Greenland interstadials. These features agree well with the palaeo record¹¹.

The main difference from previous model studies is that we do not attempt to explain a cooling in Greenland as a consequence of a Heinrich event; instead, we have assumed that the Heinrich event occurs when the climate is already in a cold stadial state, and we show that a Heinrich event in this case leads to no further cooling in Greenland.

Stability of glacial versus modern climate

Previous attempts to model rapid climate changes have started from the present-day Atlantic circulation state, and have focused on modelling cold events as a reduction or collapse of NADW formation triggered by episodic freshwater inflow into the Atlantic^{38–41}. This explanation is problematic for glacial conditions, when the climate system stayed for many thousands of years in a cold (stadial) state during periods of accumulation, not melting, of ice sheets. Moreover, improved dating accuracy has made it clear that Heinrich events (associated with large freshwater input into the North

Atlantic) occurred after major cooling events in Greenland and could therefore not be their cause.

The results of our CLIMBER-2 simulations suggest that the cold stadial periods are the normal mode of operation of the glacial climate system, representing the only stable mode of the glacial Atlantic Ocean circulation: the ‘cold’ conveyor mode with NADW formation south of Iceland. This realization shifts the focus from modelling ‘cold events’ to understanding ‘warm events’.

Our simulations show that warm events—which are similar to the observed D/O events in time evolution, amplitude and spatial pattern—can be triggered in the model as a temporary flip to the ‘warm’ conveyor mode with NADW formation in the Nordic Seas. A low-amplitude cycle in freshwater forcing is sufficient to trigger these events. What causes this cycle is not considered here—it could, for example, be ultimately due to solar variability⁴². We merely wish to demonstrate how a weak climate cycle can trigger large-amplitude episodic warm events due to the nonlinear threshold response of the Atlantic Ocean circulation.

Heinrich events, simulated in the model as a large freshwater release during a stadial, lead to a collapse of NADW formation but no major further cooling in Greenland, in agreement with the observed record. Both Heinrich and D/O events are inherently transient as the respective Atlantic circulation modes, the ‘off’ and the ‘warm’ modes, are unstable under glacial conditions.

For present-day climate the situation is the reverse, with the ‘warm’ and the ‘off’ modes representing the two stable Atlantic circulation states. Our stability analysis thus suggests that a warm climate favours the ‘warm’ conveyor belt mode (deep-water formation in the Nordic Seas), and a cold climate favours the ‘cold’ conveyor belt mode (deep-water formation south of Iceland). In the ‘cold’ mode, transient ‘flips’ to the ‘warm’ mode are easily triggered. The exact threshold value depends on the model details, but the crucial fact that it is very small is robust and underpinned by theoretical considerations. We also expect that the real climate will not have one fixed threshold value, but that this varies with the background climate. It is a robust result of our sensitivity study that the glacial ‘cold’ mode becomes more stable the colder the conditions get, that is, it gets harder to trigger ‘flips’ to the ‘warm’ mode. This could explain why the conveyor belt stays predominantly in the ‘cold’ mode during the coldest parts of the glacial⁴³, while D/O events occur almost each 1,500-year cycle during more moderate glacial conditions (50–30 kyr BP). As the climate ranges between interglacial (stable warm Atlantic mode) and full glacial (stable cold Atlantic mode) conditions, it passes through a parameter range where it ‘flickers’ between these two modes.

Once the system is in the ‘warm’ mode with convection in latitudes north of Iceland, it becomes insensitive to the applied, weak 1,500-year forcing cycle (this experiment was performed but is not detailed here). The freshwater budget of the Nordic Seas is then dominated by the vigorous circulation; anomalies in surface forcing cannot accumulate to create noticeable salinity anomalies as in the stratified ‘cold’ mode. For this reason, the Holocene climate in our model is stable with respect to the 1,500-year forcing cycle, while the glacial climate is not. We can thus explain the large fluctuations of Greenland temperature during the glacial climate in terms of ocean circulation instability, requiring only a weak trigger but not necessarily any major ice-sheet instability. In the Holocene, the 1,500-year cycle is still present⁴⁴ but is not amplified by ocean circulation instability, so that its signature is only weak. □

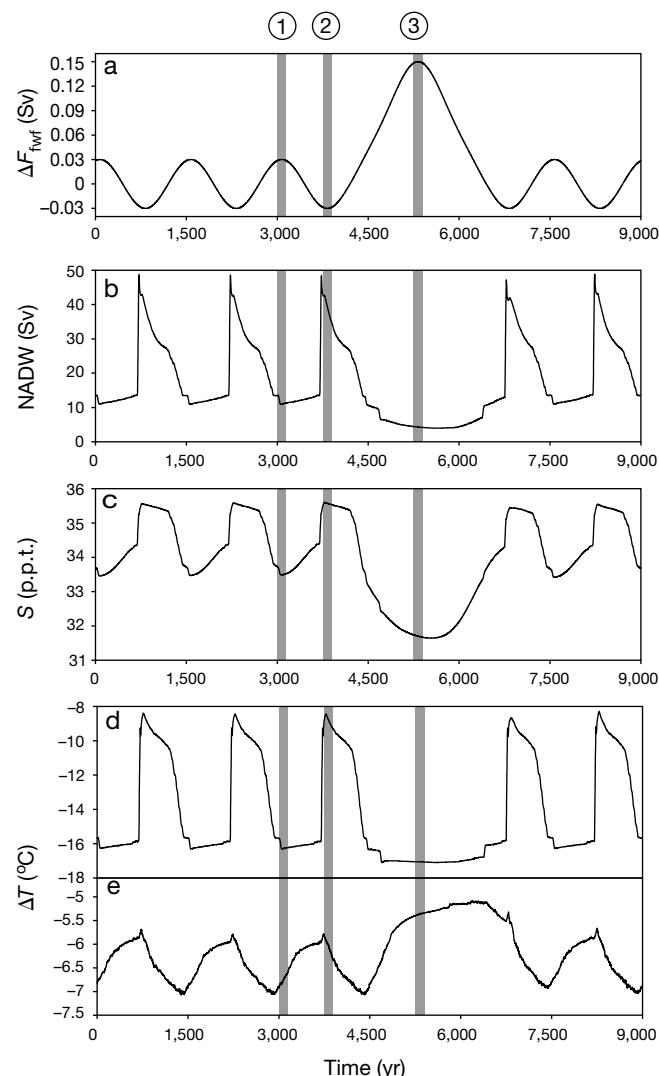


Figure 5 Simulated D/O and Heinrich events. **a**, Forcing, **b**, Atlantic overturning, **c**, Atlantic salinity (S) at 60°N , **d**, air temperature in the northern North Atlantic sector ($60\text{--}70^\circ\text{N}$), and **e**, temperature over Antarctica (temperature values are given as the difference from the present-day climate, ΔT). The vertical bars denote the times for which the difference plot is shown in Fig. 3b, c. The time sequence from model year 3,000 to 7,500 of this experiment can be viewed as a movie (see Supplementary Information).

Methods

Arctic freshwater budget

At present about 0.1 Sv of fresh water (about half of the total Arctic freshwater balance) escapes from the Arctic⁴⁵ via the Canadian archipelago and the East Greenland current (in the form of low-salinity surface currents and sea-ice transport) without mixing with the incoming salty Atlantic water. This is an important mechanism which stabilizes convection in the Nordic Seas. This process cannot be explicitly described in a zonally averaged

ocean model, and this leads to an unrealistically low threshold for convective instability of the thermohaline circulation. We therefore developed a parametrization of the freshwater bypass which is included in the new version of the CLIMBER-2 model (in contrast to the version used in ref. 15—we note that the glacial maximum simulation results do not depend on this). The parametrization has two components.

Freshwater bypass by currents

We apply an additional northward mass transport in the upper layers of the North Atlantic between 50° N and 70° N (maximum at 60° N) which mimics the gyre circulation in this region. The same amount of water is returned with a salinity equal to the average salinity between 60° N and 70° N. Thus this parametrization produces a zero net meridional mass transport but a non-zero meridional salinity transport, proportional to the local salinity gradient. The value of the gyre mass transport was tuned to obtain the corresponding observed present-day freshwater transport (0.08 Sv). For glacial conditions, this mass transport was reduced by 50% to take into account the closing of the Canadian archipelago by ice sheets and the partial closure and shallowing of the Denmark Strait. Only the quantitative details, but not the main conclusions, of this work depend on this choice.

Sea-ice export

Our parametrization of ice export is based on the assumption that the export of sea ice from the Nordic seas is proportional to the amount of sea ice in each grid cell and to the meridional wind stress. Exported ice is taken from each grid cell between 60° N and 80° N and is added to the grid cell located 1,000 km further south. Thus the total ice volume is preserved, but a southward sea-ice transport arises which adds to the freshwater transport. For modern climatic conditions this sea-ice export from the Nordic Seas is relatively small (about 0.03 Sv), but it increases during glacial conditions when the Nordic Seas were covered by ice during most of the year.

Received 5 June; accepted 13 November 2000.

1. Heinrich, H. Origin and consequences of cyclic ice rafting in the northeast Atlantic Ocean during the past 130,000 years. *Quat. Res.* **29**, 143–152 (1988).
2. Dansgaard, W. *et al.* A new Greenland deep ice core. *Science* **218**, 1273–1277 (1982).
3. Dansgaard, W. *et al.* Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364**, 218–220 (1993).
4. Clark, P. U., Webb, R. S. & Keigwin, L. D. *Mechanisms of Global Climate Change at Millennial Time Scales* (American Geophysical Union, Washington DC, 1999).
5. Alley, R. B. & Clark, P. U. The deglaciation of the Northern Hemisphere: a global perspective. *Annu. Rev. Earth Planet. Sci.* **27**, 149–182 (1999).
6. Stocker, T. Past and future reorganizations in the climate system. *Quat. Sci. Rev.* **19**, 301–319 (2000).
7. Stocker, T. F. Abrupt climate changes: from the past to the future—a review. *Int. J. Earth Sci.* **88**, 365–374 (1999).
8. Stuiver, M. & Grootes, P. M. GISP2 oxygen isotope ratios. *Quat. Res.* **53**, 277–284 (2000).
9. Alley, R. B., Anandakrishnan, S. & Jung, P. Stochastic resonance in the North Atlantic. *Paleoceanography* (in the press).
10. Sowers, T. & Bender, M. Climate records covering the last deglaciation. *Science* **269**, 210–214 (1995).
11. Blunier, T. *et al.* Asynchrony of Antarctic and Greenland climate change during the last glacial period. *Nature* **394**, 739–743 (1998).
12. Stocker, T. F. The seesaw effect. *Science* **282**, 61–62 (1998).
13. Sarnthein, M. *et al.* Changes in east Atlantic deepwater circulation over the last 30,000 years: Eight time slice reconstructions. *Paleoceanography* **9**, 209–267 (1994).
14. Keigwin, L. D., Curry, W. B., Lehman, S. J. & Johnsen, S. The role of the deep ocean in North Atlantic climate change between 70 and 130 kyr ago. *Nature* **371**, 323–326 (1994).
15. Ganopolski, A., Rahmstorf, S., Petoukhov, V. & Claussen, M. Simulation of modern and glacial climates with a coupled global model of intermediate complexity. *Nature* **391**, 350–356 (1998).
16. Petoukhov, V. *et al.* CLIMBER-2: A climate system model of intermediate complexity. Part I: Model description and performance for present climate. *Clim. Dyn.* **16**, 1–17 (2000).
17. Ganopolski, A. *et al.* CLIMBER-2: A climate system model of intermediate complexity. Part II: Model sensitivity. *Clim. Dyn.* (in the press).
18. Stocker, T. F. & Wright, D. G. A zonally averaged ocean model for the thermohaline circulation. Part II: Inter-ocean circulation in the Pacific Atlantic basin system. *J. Phys. Oceanogr.* **21**, 1725–1739 (1991).
19. Rahmstorf, S. & Ganopolski, A. Long-term global warming scenarios computed with an efficient coupled climate model. *Clim. Change* **43**, 353–367 (1999).

20. Manabe, S. & Stouffer, R. J. Two stable equilibria of a coupled ocean-atmosphere model. *J. Clim.* **1**, 841–866 (1988).
21. Kageyama, M. *et al.* The Last Glacial Maximum climate over Europe and western Siberia: a PMIP comparison between models and data. *Clim. Dyn.* **17**, 23–43 (2001).
22. Ganopolski, A., Kubatzki, C., Claussen, M., Brovkin, V. & Petoukhov, V. The role of vegetation-atmosphere-ocean interaction for the climate system during the mid-Holocene. *Science* **280**, 1916–1919 (1998).
23. Kubatzki, C., Montoya, M., Rahmstorf, S., Ganopolski, A. & Claussen, M. Comparison of a coupled global model of intermediate complexity and AOGCM for the last interglacial. *Clim. Dyn.* **16**, 799–814 (2000).
24. Claussen, M. *et al.* Simulation of an abrupt change in Saharan vegetation in the mid-Holocene. *Geophys. Res. Lett.* **26**, 2037–2040 (1999).
25. Brovkin, V., Ganopolski, A. & Svirezhev, Y. A continuous climate-vegetation classification for use in climate-biosphere studies. *Ecol. Model.* **101**, 251–261 (1997).
26. Claussen, M., Brovkin, V., Kubatzki, C., Ganopolski, A. & Petoukhov, V. Modelling global terrestrial vegetation-climate interaction. *Phil. Trans. R. Soc. Lond. B* **353**, 53–63 (1998).
27. Rahmstorf, S. Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle. *Nature* **378**, 145–149 (1995).
28. Rahmstorf, S. On the freshwater forcing and transport of the Atlantic thermohaline circulation. *Clim. Dyn.* **12**, 799–811 (1996).
29. Alley, R. B., Clark, P. U., Keigwin, L. D. & Webb, R. S. in *Mechanisms of Global Climate Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L. D.) 385–394 (American Geophysical Union, Washington DC, 1999).
30. Yiou, P. *et al.* Paleoclimatic variability inferred from the spectral analysis of Greenland and Antarctic ice-core data. *J. Geophys. Res.* **102**, 26441–26454 (1997).
31. Grootes, P. M. & Stuiver, M. Oxygen 18/16 variability in Greenland snow and ice with 10³- to 10⁵-year time resolution. *J. Geophys. Res.* **102**, 26455–26470 (1997).
32. Kreveld, S. V. *et al.* Potential links between surging ice sheets, circulation changes, and the Dansgaard-Oeschger cycles in the Irminger Sea, 60–18 kyr. *Paleoceanography* **15**, 425–442 (2000).
33. Clark, P. U., Alley, R. B. & Pollard, D. Northern hemisphere ice sheet influences on global climate change. *Science* **286**, 1104–1111 (1999).
34. Paillard, D. & Cortijo, E. A simulation of the Atlantic meridional circulation during Heinrich event 4 using reconstructed sea surface temperatures and salinities. *Paleoceanography* **14**, 716–724 (1999).
35. Cacho, I. *et al.* Dausgaard-Oeschger and Heinrich event imprints in the Alboran Sea paleotemperatures. *Paleoceanography* **14**, 698–705 (1999).
36. Bard, E., Rostek, F., Turon, J.-L. & Gendreau, S. Hydrological impact of Heinrich events in the subtropical Northeast Atlantic. *Science* **289**, 1321–1324 (2000).
37. Stocker, T. F. & Marchal, O. Abrupt climate change in the computer: is it real? *Proc. Nat. Acad. Sci. USA* **97**, 1362–1365 (2000).
38. Stocker, T. F. & Wright, D. G. Rapid transitions of the ocean's deep circulation induced by changes in surface water fluxes. *Nature* **351**, 729–732 (1991).
39. Manabe, S. & Stouffer, R. Coupled ocean-atmosphere model response to freshwater input: Comparison to Younger Dryas event. *Paleoceanography* **12**, 321–336 (1997).
40. Fanning, A. F. & Weaver, A. J. Temporal-geographical meltwater influences on the North Atlantic conveyor: implications for the Younger Dryas. *Paleoceanography* **12**, 307–320 (1997).
41. Marchal, O., Stocker, T. F. & Joos, F. in *Mechanisms of Global Climate Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L. D.) 263–284 (American Geophysical Union, Washington DC, 1999).
42. Van Geel, B. *et al.* The role of solar forcing upon climate change. *Quat. Sci. Rev.* **18**, 331–338 (1999).
43. McManus, J. F., Oppo, D. W. & Cullen, J. L. A 0.5-million-year record of millennial-scale climate variability in the North Atlantic. *Science* **283**, 971–975 (1999).
44. Bond, G. *et al.* A pervasive millennial-scale cycle in North Atlantic Holocene and glacial climates. *Science* **278**, 1257–1266 (1997).
45. Aagaard, K. & Carmack, E. C. The role of sea ice and other fresh water in the Arctic circulation. *J. Geophys. Res.* **94**, 14485–14498 (1989).
46. Grootes, P. M., Stuiver, M., White, J. W. C., Johnsen, S. & Jouzel, J. Comparison of oxygen isotope records from the GISP2 and GRIP Greenland ice cores. *Nature* **366**, 552–554 (1993).
47. Dahl-Jensen, D. *et al.* Past temperatures directly from the Greenland Ice Sheet. *Science* **282**, 268–279 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or from the London editorial office of Nature.

Correspondence and requests for materials should be addressed to A.G. (e-mail: ganopolski@pik-potsdam.de).

Internal structure of a cold dark molecular cloud inferred from the extinction of background starlight

João F. Alves*, Charles J. Lada† & Elizabeth A. Lada‡

* European Southern Observatory, Karl-Schwarzschild Straße 2, D-85748 Garching b. München, Germany

† Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

‡ Astronomy Department, University of Florida, Gainesville, Florida 32608, USA

Stars and planets form within dark molecular clouds, but little is understood about the internal structure of these clouds, and consequently about the initial conditions that give rise to star and planet formation. The clouds are primarily composed of molecular hydrogen, which is virtually inaccessible to direct observation. But the clouds also contain dust, which is well mixed with the gas and which has well understood effects on the transmission of light. Here we use sensitive near-infrared measurements of the light from background stars as it is absorbed and scattered by trace amounts of dust to probe the internal structure of the dark cloud Barnard 68 with unprecedented detail. We find the cloud's density structure to be very well described by the equations for a pressure-confined, self-gravitating isothermal sphere that is critically stable according to the Bonnor–Ebert criteria^{1,2}. As a result we can precisely specify the physical conditions inside a dark cloud on the verge of collapse to form a star.

Molecular clouds are primarily composed of molecular hydrogen mixed with trace impurities including interstellar dust grains and rare organic and inorganic molecules. Because of its symmetric structure the hydrogen molecule possesses no dipole moment and cannot produce a readily detectable signal under the conditions that characterize cold, dark clouds. The traditional methods used to derive the basic physical properties of such molecular clouds therefore make use of observations of trace H₂ surrogates, namely those rare molecules with sufficient dipole moments to be easily detected by radio spectroscopic techniques, and interstellar dust, whose thermal emission can be detected by radio continuum techniques. However, the interpretation of results derived from these methods is not always straightforward^{3,4}. Several poorly constrained effects inherent in these techniques (such as deviations from local thermodynamic equilibrium, opacity variations, chemical evolution, small-scale structure, depletion of molecules, unknown emissivity properties of the dust, unknown dust temperature) make the construction of an unambiguous picture of the physical structure of these objects a very difficult task. The deployment of sensitive, large-format infrared array cameras on large telescopes, however, has altered this situation by enabling the direct measurement of the dust extinction (that is, the overall diminution of starlight due to the combined effects of absorption and scattering by dust) toward thousands of individual background stars observed through a molecular cloud. Such measurements are free from the complications that plague molecular-line or dust emission data and enable detailed maps of cloud structure to be constructed^{5–7}.

Recently we performed very sensitive near-infrared imaging observations to map the structure of a type of dark cloud known as a Bok globule^{8,9}, one of the least complicated configurations of molecular gas that is known to form stars. The target cloud for our study, Barnard 68, is itself one of the finest examples of a Bok globule, and was selected because it is a nearby, relatively isolated and morphologically simple molecular cloud with distinct boundaries, a known distance (125 pc; ref. 10), and temperature (16 K;

ref. 11). Barnard 68 appears to be part of a string of dense clouds (Barnard 68, 69 and 70) located within a well known superbubble of hot gas, Loop I, created by the combined action of stellar winds and supernova explosions^{12,13} from extinct massive stars which were once part of the Scorpio–Centaurus OB association. This cloud lies in the direction of the centre of the Galaxy but above the Galactic



Figure 1 Visible and near-infrared images of Barnard 68. Top, deep B,V,I band (0.44 μm , 0.55 μm , 0.90 μm) image ($\sim 7' \times 7'$) of the dark molecular cloud Barnard 68 taken with ESO's Very Large Telescope (VLT) located in the Chilean Andes. The cloud is seen in projection against the Galactic bulge. At these optical wavelengths the cloud is completely opaque owing to extinction of background starlight caused by small interstellar dust particles that permeate the cloud. The complete absence of foreground stars projected onto the cloud is a result of the proximity of the cloud to the Solar System (125 pc). The outer radius of the cloud is comparable to the inner size of the Oort cloud of comets that surround the Sun ($\sim 10^4$ AU). The mass of the cloud is about twice that of the Sun. Bottom, deep B,I,K band image of the cloud constructed by combining an infrared K band (2.2 μm wavelength) image with the B and I images. The K band image was obtained with ESO's New Technology Telescope (NTT) in the Chilean Andes. At near-infrared wavelengths the cloud becomes transparent and the stars located behind the cloud clearly appear in the image. Because these stars are observed only in the longest of the three wavelength bands, they appear very red in this three-colour image. These are the stars that provide measurements of dust extinction directly through the cloud.

plane where it is projected against the rich star field of the Galactic bulge. This makes Barnard 68 an ideal candidate for an infrared extinction study for the following four reasons. First, the background bulge stars are primarily late-type (giant) stars whose intrinsic infrared colours span a narrow range and can be accurately determined from observations of nearby control fields. Second, the background star field is sufficiently rich to permit a detailed sampling of the extinction across the entire extent of the cloud. Third, the cloud is sufficiently nearby that foreground star contamination is negligible. Fourth, Barnard 68 does not appear to harbour any star-formation activity¹⁴, making it more likely that the physical conditions that characterize it reflect the initial conditions for star formation.

We used the SOFI¹⁵ near-infrared camera on the European Southern Observatory's (ESO) New Technology Telescope (NTT) to obtain deep-infrared J band (1.25 μm), H band (1.65 μm) and K band (2.16 μm) images of the cloud over two nights in March 1999. Complementary optical data were obtained with ESO's Very Large Telescope (VLT) on Cerro Paranal, fitted with the FORS1 (ref. 16) charge-coupled device (CCD) camera, during one night of March 1999. The results of the optical and near-infrared imaging are displayed in Fig. 1 top and bottom. At optical wavelengths obscuring dust within the cloud renders it opaque and completely void of stars (Fig. 1 top). However, owing to the wavelength dependence of dust extinction (that is, opacity), the cloud is essentially transparent at infrared wavelengths, enabling otherwise invisible stars behind the cloud to be imaged (Fig. 1 bottom). We detected 3,708 stars simultaneously in the deep H and K band images out of which $\sim 1,000$ stars, lying behind the cloud, are not visible at optical wavelengths. Because dust opacity decreases sharply with wavelength,

the colours of stars that are detected through a dust screen appear reddened in comparison to their intrinsic colours (for example, Fig. 1 bottom). Because the amount of reddening is directly proportional to the total extinction, we can determine the line-of-sight extinction to each star using a standard reddening law for interstellar dust and knowledge of the star's intrinsic ($H - K$) colour^{5,6}. We determined very accurate individual line-of-sight extinction measurements for all the 3,708 stars that we detected by adopting: (1) the average colour of bulge stars observed in a nearby unreddened control field as the intrinsic colour for all stars in our target field; and (2) a standard interstellar reddening law¹⁷. In this manner we have accurately sampled the dust extinction and column density distribution through the Barnard 68 cloud at more than 1,000 positions with extraordinary (pencil beam) angular resolution. Although the individual measurements are characterized by high angular resolution, our mapping of the dust column density in the cloud is highly undersampled. Consequently, we smoothed these data to construct an azimuthally averaged radial extinction (dust column density) profile of the cloud and present the result in Fig. 2. To our knowledge, this is the most finely sampled and highest signal-to-noise radial column density profile ever obtained for a dense molecular cloud. The extinction profile in Fig. 2 is the two-dimensional projection of the cloud volume density profile function, $\rho(r)$ where r is the radial distance from the centre of the cloud, and therefore provides an excellent map of the internal structure of this dense dark cloud.

As early as 1948 Bok¹⁸ pointed out that roughly spherical homogeneous-looking clouds, such as Barnard 68, resemble single dynamical units much like the polytropic models^{19,20} used to describe stellar structure. Can Barnard 68 be described as a self-gravitating, polytropic sphere of molecular gas? To investigate the physical structure of the cloud we begin with the assumptions of an isothermal equation of state and spherical symmetry. The fluid equation that describes a self-gravitating, isothermal sphere in hydrostatic equilibrium is the following well known variant of the Lane–Emden equation²¹.

$$\frac{1}{\xi^2} \frac{d}{d\xi} \left(\xi^2 \frac{d\psi}{d\xi} \right) = e^{-\psi} \quad (1)$$

where $\xi = (r/a)\sqrt{4\pi G\rho_c}$ is the non-dimensional radial parameter, a is the isothermal sound speed ($a = \sqrt{kT/m}$), ρ_c is the volume density at the origin, and $\psi(\xi) = -\ln(\rho/\rho_c)$. Equation (1) is Poisson's equation in dimensionless form for the gravitational potential, ψ , when the volume density for an isothermal gas is given by the barometric formula to be proportional to the $e^{-\psi}$, and it can be solved by numerical integration subject to the usual boundary conditions:

$$\psi(0) = 0 \quad (\text{that is, } \rho = \rho_c \text{ at } r = 0)$$

$$\text{and } \frac{d\psi(0)}{d\xi} = 0 \quad \left(\text{that is, } \frac{d\rho}{dr} = 0 \text{ at } r = 0 \right)$$

For an isothermal sphere bounded by a fixed external pressure there is a family of solutions characterized by a single parameter:

$$\xi_{\max} = \frac{R}{a} \sqrt{4\pi G\rho_c} \quad (2)$$

Here $\xi_{\max}$ is the value of ξ at the outer boundary, R (refs 1, 2). Each of these solutions corresponds to a unique cloud mass density profile. For $\xi_{\max} > 6.5$ such a gaseous configuration would be unstable to gravitational collapse². The high quality of our extinction data permits a detailed comparison with the Bonnor–Ebert predictions and we find that there is a particular solution, $\xi_{\max} = 6.9 (\pm 0.2)$, that fits the data extraordinarily well as seen in Fig. 2. This solution corresponds to a centre-to-edge density contrast of $\rho_c/\rho_R = 16.5$, which is slightly in excess of the critical contrast (14.0) predicted by the Bonnor–Ebert theory. For the

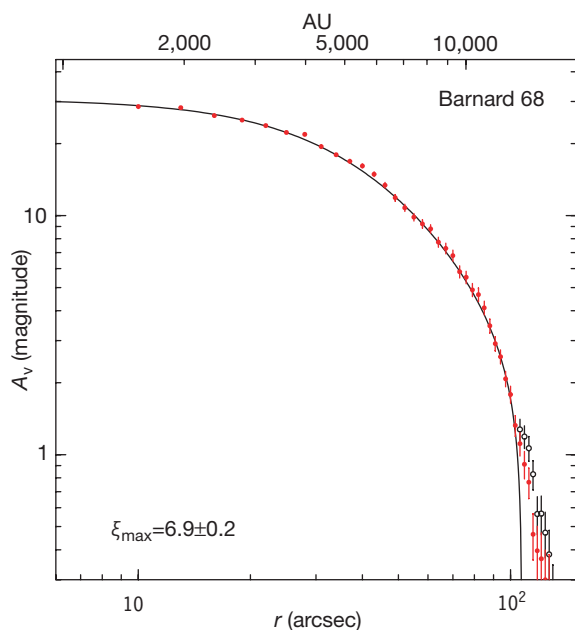


Figure 2 Azimuthally averaged radial dust column density profile of Barnard 68. By convention the dust column density is expressed in terms of magnitudes of visual extinction, A_v . The red circles show the data points for the averaged profile of a subsample of the data that do not include the cloud's southeast prominence, seen in Fig. 1. The open circles include this prominence. The error bars were computed as the r.m.s. dispersion of the extinction measurements in each averaging annulus and are smaller than the data points for the central regions of the cloud. The solid line represents the best fit of a theoretical Bonnor–Ebert sphere to the data. The close match of the data with theory indicates that the internal structure of the cloud is well characterized by the equations for a self-gravitating, pressure-confined, isothermal sphere and thus Barnard 68 seems to be a distinct dynamical unit near a state of hydrostatic equilibrium, with gravity balanced by thermal pressure.

known distance (125 pc), and temperature (16 K), Barnard 68 has a physical radius of 12,500 AU, a mass of 2.1 solar masses, and a pressure at its boundary of $P = 2.5 \times 10^{-12}$ Pa. This surface pressure is an order of magnitude higher than that of the general interstellar medium²² but it is in rough agreement with the pressure inferred for the Loop I superbubble from X-ray observations with the ROSAT satellite¹³. The close correspondence of the observed extinction profile with that predicted for a Bonnor–Ebert sphere strongly suggests that Barnard 68 is indeed an isothermal, pressure-confined and self-gravitating cloud. It is also likely to be in a state near hydrostatic equilibrium with thermal pressure primarily supporting the cloud against gravitational contraction.

For Barnard 68, $\xi_{\max}$ is very near and slightly in excess of the critical radial parameter and the cloud may be only marginally stable and on the verge of collapse. However, in the likely case that the cloud contains a static magnetic field^{23–25}, the additional internal magnetic pressure can act to stabilize it at a density contrast slightly higher than that predicted by the Bonnor–Ebert theory. Nevertheless, inevitable physical processes, such as cloud cooling, the natural reduction of internal magnetic flux by ambipolar diffusion^{26,27}, as well as any increase in external pressure, will readily destabilize the Barnard 68 cloud and result in the formation of a low-mass star, similar to the Sun.

Finally, we suggest that Barnard 68, and its neighbouring globules B69, B70 and B72 may be the precursor of an isolated and sparsely populated association of young low-mass stars similar to the recently identified TW Hydra²⁸ association. The TW Hydra association is a stellar group near the solar system consisting of a handful of young low-mass, Sun-like stars. The existence of such a young stellar group presents an interesting problem to astronomers because its origin is difficult to explain given its youth and relatively large distance from known sites of star formation. Bok globules such as those in the Barnard 68 group are thought to be remnant dense cores produced as a result of the interaction of massive O stars and molecular clouds²⁹. Over their short lifetimes such massive stars, through ionization, stellar winds and ultimately supernova explosions, very effectively disrupt the molecular clouds from which they formed. In the process large shells of expanding gas are created. When surrounding clouds are disrupted by the passage of these shells, a few of their most resilient dense cores will be left behind, embedded within the shell's hot interior. Remnant cores with just the right mass can establish pressure equilibrium with the hot gas within the shell and survive to become Bok globules. Eventually, as a result of the processes described above, these clouds will evolve to form low-mass, Sun-like stars which are relatively isolated and far from the original birthplace of the O stars. Up to 35% of all Bok globules contain newly formed stars³⁰ and thus it is likely that our observations of the starless Barnard 68 cloud provide the first detailed description of the initial conditions which exist before the collapse of dark globules and the formation of isolated, low-mass stars. □

Received 19 September; accepted 3 November 2000.

1. Ebert, R. Über die Verdichtung von HI-Gebieten. *Z. Astrophys.* **37**, 217–232 (1955).
2. Bonnor, W. Boyle's Law and gravitational instability. *Mon. Not. R. Astron. Soc.* **116**, 351–359 (1956).
3. Alves, J., Lada, C. & Lada, E. Correlation between gas and dust in molecular clouds: L977. *Astrophys. J.* **515**, 265–274 (1999).
4. Chandler, C. & Richer, J. The structure of protostellar envelopes derived from submillimeter continuum images. *Astrophys. J.* **350**, 851–866 (2000).
5. Lada, C., Lada, E., Bally, J. & Clemens, D. Dust extinction and molecular gas in the dark cloud IC 5146. *Astrophys. J.* **429**, 694–709 (1994).
6. Alves, J., Lada, C. J., Lada, E. A., Kenyon, S. J. & Phelps, R. Dust extinction and molecular cloud structure: L 977. *Astrophys. J.* **506**, 292–305 (1998).
7. Lada, C., Alves, J. & Lada, E. Infrared extinction and the structure of the IC 5146 dark cloud. *Astrophys. J.* **512**, 250–259 (1999).
8. Bok, B. & Reilly, E. Small dark nebulae. *Astrophys. J.* **105**, 255–257 (1947).
9. Clemens, D. & Barvainis, R. A catalog of small, optically selected molecular clouds: optical, infrared, and millimeter properties. *Astrophys. J. Suppl. Ser.* **68**, 257–286 (1988).
10. de Geus, E., de Zeeuw, P. & Lub, J. Physical parameters of stars in the Scorpius–Centaurus OB association. *Astron. Astrophys.* **216**, 44–61 (1989).
11. Bourke, T., Hyland, A., Robinson, G., James, S. & Wright, C. Studies of star formation in isolated small

- dark clouds—II A southern ammonia survey. *Mon. Not. R. Astron. Soc.* **276**, 1067–1084 (1995).
12. Quigley, M. & Haslam, C. Structure of the radio continuum background at high galactic latitudes. *Nature* **208**, 741–743 (1965).
13. Breitschwerdt, D., Freyberg, M. & Egger, R. Origin of HI clouds in the local bubble. *Astron. Astrophys.* **361**, 303–320 (2000).
14. Reipurth, B., Nyman, L. & Chini, R. Protostellar candidates in southern molecular clouds. *Astron. Astrophys.* **314**, 258–264 (1996).
15. Moorwood, A., Cuby, J. & Lidman, C. SOFI sees first light at the NTT. *ESO Messenger* **91**, 9–13 (1998).
16. Appenzeller, I. et al. Successful commissioning of FORSI—the first optical instrument on the VLT. *ESO Messenger* **94**, 1–6 (1998).
17. Mathis, J. Interstellar dust and extinction. *Annu. Rev. Astron. Astrophys.* **28**, 37–70 (1990).
18. Bok, B. *Centennial Symposia* (Harvard Observatory Monographs No. 7, Harvard-College Observatory, Cambridge, 1948).
19. Lane, J. *Am. J. Sci. Arts, Series 2* **4**, 57 (1870).
20. Emden, R. *Gaskugeln* (Teubner, Leipzig, 1907).
21. Chandrasekhar, S. in *An Introduction to the Study of Stellar Structure* 156 (Dover, Toronto, 1967).
22. McKee, C. in *The Origin of Stars and Planetary Systems* (eds Lada, C. & Kylafis, N.) 29–66 (Kluwer, Dordrecht, 1999).
23. Nakano, T. Quasistatic contraction of magnetic protostars due to magnetic flux leakage—Part One—formulation and an example. *Publ. Astron. Soc. Jpn* **31**, 697–712 (1979).
24. Lizano, S. & Shu, F. Molecular cloud cores and bimodal star formation. *Astrophys. J.* **342**, 834–854 (1989).
25. Basu, S. & Mouschovias, T. Magnetic braking, ambipolar diffusion, and the formation of cloud cores and protostars I—Axisymmetric solutions. *Astrophys. J.* **432**, 720–741 (1994).
26. Shu, F., Allen, A., Shang, H., Ostriker, E. & Li, Z. in *The Origin of Stars and Planetary Systems* (eds Lada, C. & Kylafis, N.) 193–226 (Kluwer, Dordrecht, 1999).
27. Mouschovias, T. & Ciolek, G. in *The Origin of Stars and Planetary Systems* (eds Lada, C. & Kylafis, N.) 305–340 (Kluwer, Dordrecht, 1999).
28. Rucinski, S. & Krautter, J. TW Hya: a T Tauri star far from any dark cloud. *Astron. Astrophys.* **121**, 217–225 (1983).
29. Reipurth, B. Star formation in Bok globules and low-mass clouds. I—The cometary globules in the Gum Nebula. *Astron. Astrophys.* **117**, 183–198 (1983).
30. Launhardt, R. & Henning, T. Millimetre dust emission from northern Bok globules. *Astron. Astrophys.* **326**, 329–346 (1997).

Acknowledgements

We thank M. Lombardi for fruitful discussions and assistance, the Paranal Science Operations team for observing Barnard 68 with FORS1 on Very Large Telescope (VLT) Antu, R. West and E. Janssen for composing Figure 1 top, and R. Hook and R. Fosbury for composing Figure 1 bottom. We also thank M. Petr for helpful discussions during the preparation of the VLT observations. E.A.L. acknowledges support from a Presidential Early Career Award for Scientists and Engineers to the University of Florida.

Correspondence and requests for materials should be addressed to J.A. (e-mail: jalves@eso.org).

Quantum metallicity in a two-dimensional insulator

V. Yu. Butko*† & P. W. Adams*

* Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803, USA

† Ioffe Physical Technical Institute (PTI), Russian Academy of Sciences, Polytekhnicheskaya Street, 26, 194021, St Petersburg, Russia

One of the most far-reaching problems in condensed-matter physics is to understand how interactions between electrons, and the resulting correlations, affect the electronic properties of disordered two-dimensional systems. Extensive experimental^{1–6} and theoretical^{7–11} studies have shown that interaction effects are enhanced by disorder, and that this generally results in a depletion of the density of electronic states. In the limit of strong disorder, this depletion takes the form of a complete gap^{12,13} in the density of states. It is known that this ‘Coulomb gap’ can turn a pure metal film that is highly disordered into a poorly conducting insulator¹⁴, but the properties of these insulators are not well understood. Here we investigate the electronic properties of disordered beryllium films, with the aim of disentangling the effects of the Coulomb gap and the underlying disorder. We show that the gap is suppressed by a magnetic field and that this drives

the strongly insulating beryllium films into a low-temperature 'quantum metal' phase with resistance near the quantum resistance $R_Q = h/e^2$, where h is Planck's constant and e is the electron charge.

The theory of interaction effects in disordered electronic systems has for the most part been developed for two extreme limits. In the two-dimensional (2D) weak disorder limit in which the film resistance R is much less than $R_Q = 26 \text{ k}\Omega$, electron-electron (e - e) interactions produce an anomalous logarithmic suppression of the density of states (DOS) at the Fermi energy^{9,15,16}. These interactions also, along with coherent backscattering^{4,7}, produce a weak logarithmic increase in R with decreasing temperature, which is commonly referred to as weak localization. We will refer to metal films in this limit as being 'weakly metallic' because the observed increases in R with decreasing temperature are typically of the order of a few per cent. In the opposite limit, that is, the strongly insulating regime, the e - e interactions grow rapidly and ultimately produce the Coulomb gap in the DOS¹². The 2D gap spectrum is predicted to be linear in energy¹², and, in contrast to the weak localization regime, it completely dominates the transport behaviour of the system and can produce many orders of magnitude increases in R with decreasing temperature.

We chose to fabricate our metal films using beryllium, because it is known to form a smooth, dense, non-granular morphology when thermally evaporated onto glass¹⁷. This morphology is crucial in that it ensures that the measured resistance is representative of e - e correlation effects and not grain charging effects. In the present study we used films with thicknesses of $\sim 1.7 \text{ nm}$ and corresponding low-temperature sheet resistances in the range $3R_Q$ – $5R_Q$. They were deposited by thermally evaporating 99.5% pure beryllium metal onto fire-polished glass substrates held at 84 K. The evaporations were performed in a 4×10^{-7} torr vacuum at a rate of $\sim 0.15 \text{ nm s}^{-1}$. The film area was $1.5 \text{ mm} \times 4.5 \text{ mm}$. The tunnel junctions were formed by exposing the films to the atmosphere for 0.1–3 h in order to form a native oxide. Then a 20-nm-thick Ag counter-electrode was deposited directly on top of the film, with the oxide serving as the tunnel barrier; the junction area was $0.7 \text{ mm} \times 0.7 \text{ mm}$. The resistance of all of the samples discussed below was high enough to suppress completely the superconducting phase. The film resistances were measured using a standard four-probe a.c. current-voltage (I - V) technique, and the tunnelling conductances were measured using a d.c. technique. Magnetoresistance (MR) measurements were made with the magnetic field oriented perpendicular to the film surface.

In Fig. 1 we plot the temperature dependence of the film resistance (normalized by R_Q) for two critically disordered samples;

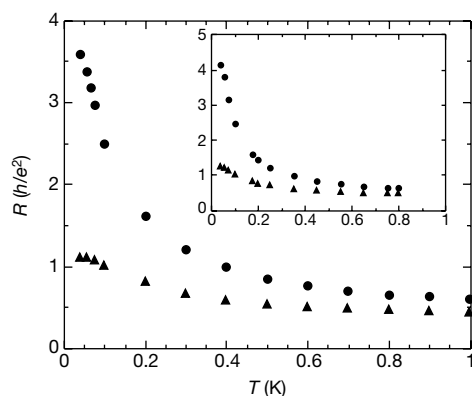


Figure 1 Temperature dependence. The figure shows the resistance of film B57 in units of h/e^2 in zero field (circles) and at a perpendicular magnetic field of $H = 8.4 \text{ T}$ (triangles) as a function of temperature. Inset, temperature dependence of the resistance of sample B55.

the resistance of each film was measured in zero field and in a field of 8.4 T. The zero-field data are strongly insulating, but the high-field data appear to be more metallic in character: the magnetic field is producing a significant suppression of the insulating phase. This is also evident in Fig. 2, where we show the low-temperature MR of the films. We note that the low-field MR is positive, but above 1 T the MR becomes strongly negative. Similar behaviour has recently been reported in much higher-resistance Be films, $R \approx 100R_Q$, which obeyed the Efros-Shklovskii hopping law^{12,14,18}, $R = (R_Q/2)\exp(T_0/T)^{1/2}$, where T_0 is the correlation energy and T is the temperature. These extremely insulating films exhibited a small, positive MR at low field, followed by a tenfold negative MR with no signs of saturation up to $\sim 9 \text{ T}$. The negative MR was well described by a $1/R \propto H$ dependence¹⁸ (where H is the magnetic

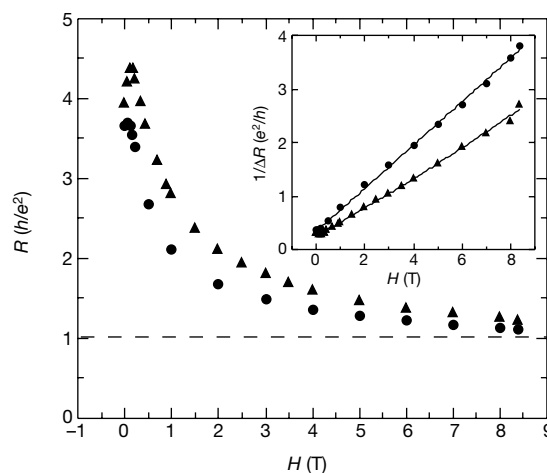


Figure 2 Low-temperature magnetoresistance. The figure shows the resistance in units of h/e^2 as a function of magnetic field H at $T = 40 \text{ mK}$, for samples B57 (circles) and B55 (triangles). We note the saturation at $R \approx h/e^2$. Inset, linear behaviour after subtracting a saturation resistance of $0.85 h/e^2$, $\Delta R = R - 0.85$. The solid lines are guides to the eye.

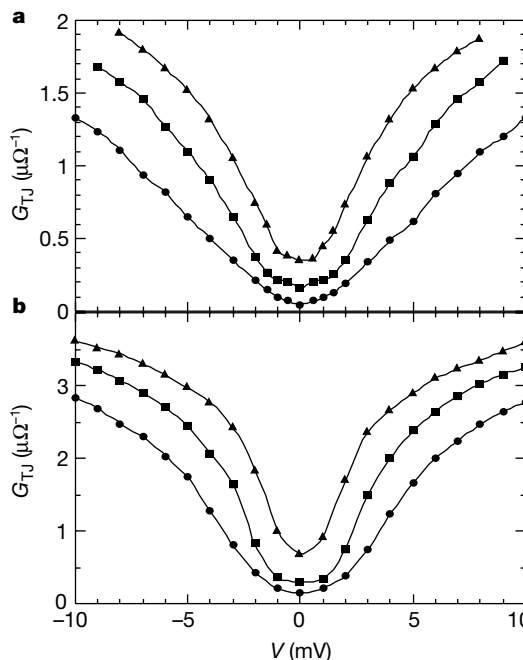


Figure 3 Tunnelling density of states. **a**, Low-temperature tunnelling conductance (G_{TJ}) of the B55 film at $H = 0$ (circles), $H = 3.0 \text{ T}$ (squares) and $H = 8.4 \text{ T}$ (triangles). **b**, As **a** but for the B57 film. The tunnelling conductance is proportional to the density of electronic states which clearly increases with increasing field. The solid lines are guides to the eye.

field intensity) which was shown to arise from a field suppression of the correlation energy T_0 , suggesting that the Coulomb gap itself is field-dependent. The overall structure of the data in Fig. 2 is consistent with the MR behaviour of the highly insulating films of ref. 18, with the notable exception of the saturation of the MR to the quantum resistance.

The saturation behaviour in Fig. 2 suggests that the film resistance can be written as the sum of two terms:

$$R = R_{ee} + \beta R_Q \quad (1)$$

where R_{ee} is a field-dependent correlation resistance and the second term is a field-independent background. The constant β is only weakly temperature-dependent, and is of order unity. If we take $\beta = 0.85$ and subtract the background resistance from the MR data, then we find that $1/R_{ee} \propto H$, as shown in the inset of Fig. 2. This is precisely the field dependence reported in the $R \approx 100R_Q$ Be films of ref. 18, suggesting that the negative MR of those films would have also saturated to the quantum resistance at $H \approx 100$ T. Indeed, a compelling interpretation of the data in Figs 1 and 2 is that the field drives the system out of an insulating phase, which is dominated by R_{ee} into a weakly metallic phase at a universal quantum resistance. We have termed this high-field phase a 'quantum metal' by virtue of its extremely weak temperature dependence and the fact that its resistance is near the quantum resistance R_Q . In terms of the MR measurements of films of ref. 18, the quantum metal phase would have been reached at a field high enough to drive T_0 to zero.

The correlation resistance R_{ee} and its field dependence manifest the physical processes that are also associated with the emergence of the Coulomb gap. In Fig. 3 we plot the tunnelling conductance as a

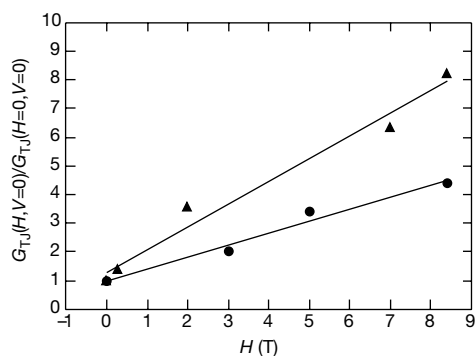


Figure 4 Zero-bias tunnelling conductance. The figure shows relative zero-bias tunnelling conductance as a function of magnetic field at $T = 100$ mK for samples B55 (triangles) and B57 (circles). The solid lines are guides to the eye.

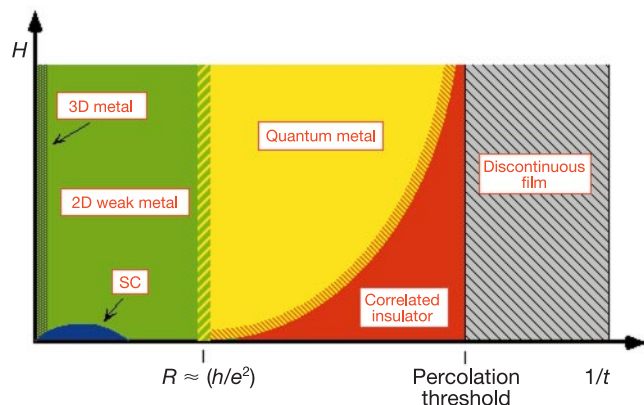


Figure 5 Phase diagram. The figure shows the low-temperature electronic phase diagram of Be films where H is the magnetic field intensity and t is the film thickness. We interpret $1/t$ as a disorder parameter.

function of bias voltage at several field values for both samples. At low temperatures, the tunnelling conductance is proportional to the electronic DOS of the films¹⁹ with $V = 0$ corresponding to the Fermi energy. We note that the zero-field curves in Fig. 3 show an almost complete depletion of states in the vicinity of $V = 0$. This is consistent with the linear Coulomb gap spectra reported in studies of much higher-resistance Be films in ref. 14. The most interesting aspect of Fig. 3, however, is that the application of field causes a significant increase in the number of states near the Fermi energy²⁰ with a corresponding decrease in film resistance.

We believe that the field dependence of the DOS spectrum is reflected in the field dependence of R_{ee} and that $R_{ee} \rightarrow 0$ in the limit of $H \rightarrow \infty$. The quantum resistance emerges once the gap is lifted by the field. We point out here that the saturation of the MR in Fig. 2 is not reflected in the DOS field dependence. In Fig. 4 we plot the relative tunnelling conductance at zero bias (that is, the Fermi energy) as a function of field. These data clearly show that the number of excess states is proportional to the field with no sign of saturation up to 9 T, though the MR does saturate in this field range.

The low-temperature magnetotransport characteristics of the Be films are most effectively summarized using the schematic phase diagram in Fig. 5. The x axis measures the amount of disorder, which we parametrize as $1/t$ where t is the film thickness. The y axis is the applied magnetic field. In the limit of $1/t \rightarrow 0$ the films are three-dimensional (3D) and have finite resistance at zero temperature. As the thickness is lowered, however, the 3D to 2D threshold is crossed, and the films become weakly metallic with a perturbative logarithmic increase in resistance with decreasing temperature and a perturbative logarithmic suppression of the DOS near the Fermi energy. We note that Be also has a locally enhanced superconducting phase in the weakly metallic regime, shown in blue in Fig. 5 (The bulk superconducting transition temperature of Be is $T_c = 26$ mK: at $t \approx 5$ nm, the film resistance is $R \approx 0.1$ k Ω and $T_c \approx 0.6$ K; at $t \approx 1.8$ nm, $R \approx 10$ k Ω and $T_c \approx 0$). As the film thickness is further decreased in zero field the resistance increases. When $R \approx R_Q$ the depletion of states near the Fermi energy grows exponentially¹⁴ with increasing R , signalling the emergence of the Coulomb gap. This crossover is depicted by the green-yellow hatched boundary between the weakly and strongly correlated regions in the diagram. Further decreases in thickness at zero field result in films with $R \gg R_Q$. These films are correlated insulators, shown in red, and are characterized by a well defined Coulomb gap and Efros-Shklovskii hopping transport. The further one moves to the right in the correlated insulator region of the diagram the greater the correlation energy T_0 becomes, until one eventually reaches the percolation threshold of the film, at which point it is no longer electrically continuous. This occurs at $t \approx 1$ nm in Be films. We have shown that the application of a magnetic field lifts the Coulomb gap, thereby producing extremely large decreases in resistance at low temperatures. As the diagram in Fig. 5 indicates, we believe that this process eventually leads to a saturation in the resistance near R_Q even for films deep in the correlated insulator regime. This weakly temperature-dependent, high-field state is depicted in yellow and essentially behaves as a weak metal. Thus the yellow region in Fig. 5 describes films of varying disorder but with resistance near R_Q . The almost ideal morphological characteristics of Be films, along with the fact that Be is intrinsically non-magnetic, lead us to believe that the general features of Fig. 5 are generic to high-density, disordered, 2D electron gases. □

Received 24 July; accepted 9 November 2000.

- Hsu, S.-Y. & Valles, J. M. Jr Electron tunneling into strongly disordered films: the influence of structure on e-e interactions. *Phys. Rev. B* **49**, 16600–16604 (1994).
- Vaknin, A., Ovadyahu, Z. & Pollak, M. Evidence for interactions in nonergodic electronic transport. *Phys. Rev. Lett.* **81**, 669–672 (1998).
- Nissim, M., Leareah, Y. & Rosenbaum, R. Dominating Coulomb-interaction effects in amorphous In_xO_y films. *Phys. Rev. B* **40**, 6351–6355 (1989).
- Bishop, D. J., Dynes, R. C. & Tsui, D. C. Magnetoresistance in Si metal-oxide-semiconductor field-effect transistors: Evidence of weak localization and correlation. *Phys. Rev. B* **26**, 773–779 (1982).

5. Das Sarma, S. & Pinczuk, A. (eds) *Perspectives in Quantum Hall Effects* (Wiley, New York, 1997).
6. Kravchenko, S. V., Kravchenko, G. V., Furneaux, J. E., Pudalov, V. M. & D'Iorio, M. Possible metal-insulator transition at $B = 0$ in two dimensions. *Phys. Rev. B* **50**, 8039–8042 (1994).
7. Lee, P. A. & Ramakrishnan, T. V. Disordered electronic systems. *Rev. Mod. Phys.* **57**, 287–337 (1985).
8. Ando, T., Fowler, A. B. & Stern, F. Electronic properties of two-dimensional systems. *Rev. Mod. Phys.* **54**, 437–672 (1982).
9. Altshuler, B. L., Aronov, A. G., Gershenson, M. E. & Sharvin, Yu. V. Quantum effects in disordered metal films. *Sov. Sci. Rev. A Phys.* **9**, 223–353 (1987).
10. Belitz, D. & Kirkpatrick, T. R. The Anderson-Mott transition. *Rev. Mod. Phys.* **66**, 261–390 (1994).
11. Pastor, A. A. & Dobrosavljevic, V. Melting of the electron gas. *Phys. Rev. Lett.* **83**, 4642–4645 (1999).
12. Shklovskii, B. I. & Efros, A. L. *Electronic Properties of Doped Semiconductors* (Springer, New York, 1984).
13. Massey, J. G. & Lee, M. Electron tunneling study of Coulomb correlations across the metal-insulator transition in Si:B. *Phys. Rev. Lett.* **77**, 3399–3402 (1996).
14. Butko, V. Yu., DiTusa, J. F. & Adams, P. W. Coulomb gap: How a metal film becomes an insulator. *Phys. Rev. Lett.* **84**, 1543–1546 (2000).
15. Gershenson, M. E., Gubankov, V. N. & Falei, M. I. Effect of electron-electron interaction on state density in 2D Al films. *Pis'ma Zh. Eksp. Teor. Fiz.* **41**, 435–439 (1985) [*JETP Lett.* **41**, 535–538 (1985)].
16. Imry, Y. & Ovadyahu, Z. Density of states anomalies in disordered metal films. *Phys. Rev. Lett.* **49**, 841–844 (1982).
17. Adams, P. W., Herron, P. & Meletis, E. I. First-order spin-paramagnetic transition and tri-critical point in ultrathin Be films. *Phys. Rev. B* **58**, 2952–2955 (1998).
18. Butko, V. Yu., DiTusa, J. F. & Adams, P. W. Tenfold magnetoconductance in a non-magnetic metal film. *Phys. Rev. Lett.* **85**, 162–165 (2000).
19. Tinkham, M. *Introduction to Superconductivity* (McGraw-Hill, New York, 1996).
20. Sood, B. R. Magnetic-field-dependent zero-bias anomaly in Al-oxide-Ga (granular) thin-film tunnel junctions. *Phys. Rev. B* **25**, 6064–6066 (1982).

Acknowledgements

We thank J. DiTusa, D. Browne, B. Shklovskii, V. Dobrosavljevic, I. Aleiner and B. Altshuler for discussions. This work was supported by the NSF.

Correspondence and requests for materials should be addressed to P.W.A. (e-mail: adams@rouge.phys.lsu.edu).

The relationship between fragility, configurational entropy and the potential energy landscape of glass-forming liquids

Srikanth Sastry

Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur Campus, Bangalore 560064, India

Glass is a microscopically disordered, solid form of matter that results when a fluid is cooled or compressed in such a manner that it does not crystallize. Almost all types of materials are capable of glass formation, including polymers, metal alloys and molten salts. Given such diversity, general principles by which different glass-forming materials can be systematically classified are invaluable. One such principle is the classification of glass-formers according to their fragility¹. Fragility measures the rapidity with which a liquid's properties (such as viscosity) change as the glassy state is approached. Although the relationship between the fragility, configurational entropy and features of the energy landscape (the complicated dependence of energy on configuration) of a glass-former have been analysed previously², a detailed understanding of the origins of fragility is lacking. Here I use simulations to analyse the relationship between fragility and quantitative measures of the energy landscape for a model liquid whose fragility depends on its bulk density. The results reveal that fragility depends on changes in the vibrational properties of individual energy minima in addition to their total number and spread in energy. A thermodynamic expression for fragility is

derived, which is in quantitative agreement with kinetic fragilities obtained from the liquid's diffusivity.

Glass-forming liquids grow increasingly viscous upon cooling, till the viscosity becomes so large that they fail to flow on experimental timescales. While remaining microscopically disordered like a liquid, they manifest mechanical properties of a solid; that is, they transform to a glass. By convention, the glass transition temperature T_g is where the viscosity reaches a value of 10^{12} Pa s. The approach to this large viscosity, however, differs from one liquid to another. When displayed in an Arrhenius plot of $\log(\text{viscosity})$ versus inverse temperature $1/T$, some liquids (such as silica) show a steady, linear increase, while others display a much steeper dependence on $1/T$ (refs 1, 3), as illustrated in Fig. 1a (inset). The former are 'strong' liquids, and the latter, 'fragile'. This range of behaviour is implicit in the Vogel–Fulcher–Tammann–Hesse (VFT) form, observed to describe the T dependence of viscosity (as well as diffusivity and relaxation times) in many glass-formers. The VFT relation may be written as:

$$\eta = \eta_0 \exp \left[\frac{1}{K_{\text{VFT}}(T/T_0 - 1)} \right] \quad (1)$$

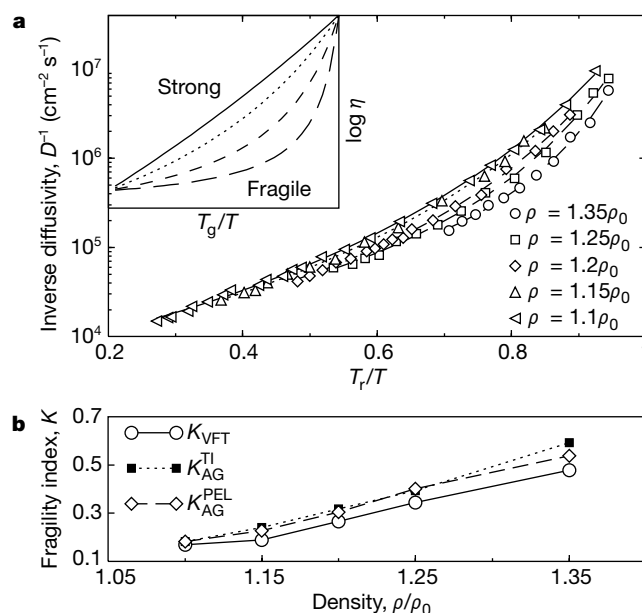


Figure 1 Fragility plot of diffusivities, and kinetic and thermodynamic fragility indices. Results shown are obtained from molecular dynamics simulations of 204 A-type and 52 B-type particles interacting via the Lennard–Jones (LJ) potential. Argon units are used for A-type particles: LJ parameter $\epsilon_{\text{AA}} = k_B \times 119.8 \text{ K}$, $\sigma_{\text{AA}} = 0.3405 \times 10^{-9} \text{ m}$, $m_{\text{A}} = 6.6337 \times 10^{-26} \text{ kg}$; and $\epsilon_{\text{AB}}/\epsilon_{\text{AA}} = 1.5$, $\epsilon_{\text{BB}}/\epsilon_{\text{AA}} = 0.5$, $\sigma_{\text{AB}}/\sigma_{\text{AA}} = 0.8$, and $\sigma_{\text{BB}}/\sigma_{\text{AA}} = 0.88$, $m_{\text{B}}/m_{\text{A}} = 1$. Densities are reported in units of $\rho_0 = 2.53 \times 10^{28} \text{ m}^{-3}$ or 1.678 g cm^{-3} . The liquid–gas critical point is located roughly at $T_c = 130 \text{ K}$, $\rho_c = 0.416\rho_0$. **a**, Diffusivities D from molecular dynamics simulations (details of the simulations are as in ref. 10) are displayed in a ‘fragility plot’ for five densities. The reference temperature T_r is chosen such that $D(T_r) = 2.44 \times 10^{-8} \text{ cm}^2 \text{s}^{-1}$, which is slightly below the lowest D values measured. The VFT extrapolation is used to locate T_r . Fitting values D_0 (preexponent) are in the range of 0.87 to $2.0 \times 10^{-4} \text{ cm}^2 \text{s}^{-1}$, close to the experimentally observed values. **b**, Fragility index K obtained (1) from VFT fits to diffusivities D ; and from the configurational entropy obtained (2) from thermodynamic integration (TI), and (3) from analysis of the potential energy landscape (PEL). Dimensionless thermodynamic fragility indices are obtained by dividing $K_{\text{AG}}^{\text{TI}}$, $K_{\text{AG}}^{\text{PEL}}$ by $N\epsilon_{\text{AA}}$. Comparison with data in ref. 11 shows that at the low-density end, the fragility of the model liquid compares with that of considerably fragile liquids such as toluene, orthoterphenyl and salol, while at the high-density end, the fragility is extremely high. Temperatures of vanishing diffusivity T_0 from VFT fits are 18.76 K, 24.98 K, 35.97 K, 47.86 K and 75.75 K for $\rho/\rho_0 = 1.1, 1.15, 1.2, 1.25$ and 1.35 respectively.

where T_0 is the temperature of apparent divergence of viscosity, and K_{VFT} is a material-specific parameter quantifying the kinetic fragility; more fragile liquids have larger K_{VFT} values. This behaviour of viscosity correlates with the jump in heat capacity at the glass transition—the more fragile the liquid, the sharper and bigger the jump. Rationalization of this correlation comes from the Adam–Gibbs relationship⁴ which predicts a dependence of viscosity on the configurational entropy S_c (described later) of the liquid:

$$\eta = \eta_0 \exp \left[\frac{A}{TS_c} \right] \quad (2)$$

which results in the VFT relation if S_c has the form $TS_c = K_{\text{AG}}(T/T_K - 1)$ (ref. 5), where T_K is the ideal glass transition or Kauzmann temperature⁴ below which $S_c = 0$. The jump in heat capacity is proportional to K_{AG} , and is larger for more fragile glass-formers.

Fragility is analysed here for a binary mixture of particles interacting via the Lennard–Jones potential, widely studied as a model glass-former^{6–10}. The T dependence of diffusivity D (to characterize dynamics) and configurational entropy are obtained for a range of bulk densities ρ . The diffusivities $D(\rho, T)$, are shown in Fig. 1a in a ‘fragility plot’: $-\log[D(\rho, T)]$ versus T_i/T , where T_i is a reference temperature (akin to T_g in the usual fragility plot) at which D at each density reaches a fixed, small value. By comparison with the inset, the liquid is seen to become more fragile as its density increases. Figure 1b shows the kinetic fragility index, K_{VFT} , obtained from VFT (equation (1)) fits to diffusivity data at each density. The density dependence of K_{VFT} quantifies the increase of fragility with density, apparent from data in Fig. 1a.

A thermodynamic explanation of fragility, as discussed above (also refs. 11, 12), lies in explaining the rapidity of change with T of the configurational entropy, S_c . In the inherent structure (IS) formalism^{13,14} used here, the configuration space of a liquid is divided into basins of local potential energy minima (‘inherent structures’). S_c at a given T derives from the multiplicity of local potential energy minima sampled by the liquid at that T , and equals the difference of the total entropy and the vibrational entropy of typical basins sampled, S_{vib} :

$$S_c(\rho, T) = S_{\text{total}}(\rho, T) - S_{\text{vib}}(\rho, T) \quad (3)$$

For all the results presented here, the basin entropy S_{vib} is calculated by approximating each basin as a harmonic well (valid at sufficiently low T)^{8–10,15}. In this approximation, the entropy of a given basin, arising from vibrational motion within the basin is given by:

$$S_{\text{vib}} = k_B \sum_{i=1}^{3N} 1 - \log \left(\frac{h\nu_i}{k_B T} \right) \quad (4)$$

where ν_i are the vibrational frequencies of the given basin, k_B is Boltzmann’s constant and h is Planck’s constant. The basin free energy is $F_{\text{vib}} = U_{\text{vib}} - TS_{\text{vib}}$, where $U_{\text{vib}} = 3Nk_B T$ is the internal energy. Apart from the explicit temperature dependence of S_{vib} above, an implicit dependence also exists because vibrational frequencies change from one basin to another, and different basins are sampled at different temperatures. From the expression above, it is clear that the entropy difference between any two basins arises solely from the difference in their vibrational frequencies, and is independent of temperature.

The configurational entropy is evaluated here using two methods, referred to as the thermodynamic integration (TI) method, and the potential energy landscape (PEL) method, respectively. In both cases, inherent structures are obtained by performing local potential energy minimization for a subset of configurations sampled by a liquid at each ρ and T . Vibrational frequencies are calculated for each inherent structure, and the basin entropy S_{vib} is obtained from equation (4). In the TI method (described in detail elsewhere^{8–10}; see

also caption of Fig. 2), the total entropy S_{total} is obtained using pressure and internal energy values from simulations. Equation (3) is used to obtain S_c . In contrast, the PEL method is based on constructing the distribution of inherent structures in order to calculate S_{total} and S_c , as described later.

Figure 2a shows the Adam–Gibbs plot, where $\log D$ is plotted against $(TS_c)^{-1}$. The Adam–Gibbs relation (equation (2)) applies well at each density, as was also observed previously for the hard sphere liquid¹⁶ and water¹⁷. Figure 2b shows (as lines) the T -dependence of TS_c (from TI). All curves are nearly linear, but with slight increases of slope at lower T . The slopes of TS_c (in the T range where diffusivities are measured), $K_{\text{AG}}^{\text{TI}}$, are used as the quantitative index of thermodynamic fragility, and are plotted in Fig. 1b. $K_{\text{AG}}^{\text{TI}}$ are nearly the same as the kinetic fragility index K_{VFT} , thereby validating the idea that the rate of increase of TS_c does indeed determine the fragility.

However, this analysis provides no insight into how the fragility relates to features of the energy landscape. To this end, I now examine the distribution of energy minima, and properties of individual basins at fixed density ρ . The probability distribution $P(\Phi, T)$ that an inherent structure of energy Φ is sampled at temperature T depends on (1) the energy Φ ; (2) the structure of

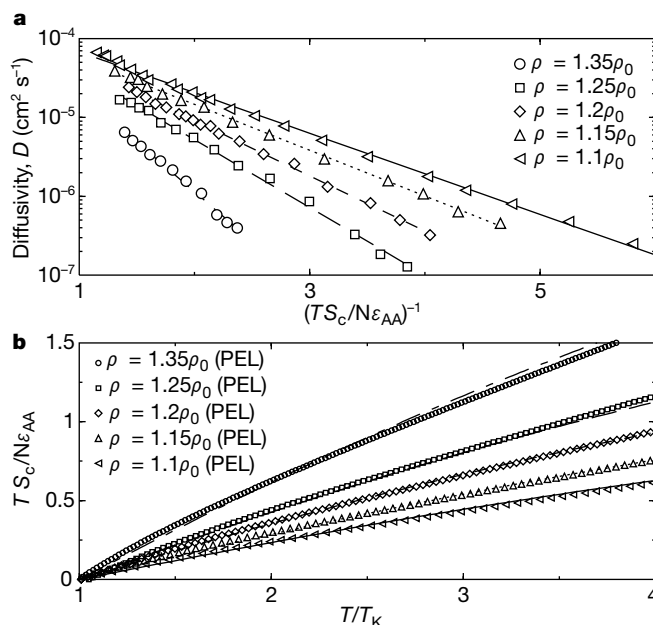


Figure 2 Adam–Gibbs plot of configurational entropy versus temperature. **a**, The Adam–Gibbs plot displaying $\log(D)$ versus $(TS_c)^{-1}$. TS_c is expressed in units of ϵ_{AA} . The Adam–Gibbs expectation of a linear dependence is well satisfied at each density. **b**, Configurational entropy obtained from thermodynamic integration (TI) (lines, not labelled), and from analysis of the potential energy landscape (points, labelled PEL), plotted against temperature T scaled to the Kauzmann temperature T_K obtained from thermodynamic integration. In the TI method, the total entropy S_{total} of the liquid is obtained from its total free energy F_{total} , from $F_{\text{total}} = U_{\text{total}} - TS_{\text{total}}$, where U_{total} is the internal energy. Thermodynamic identity $P = \rho^2((\partial F_{\text{total}}/N)/(\partial \rho)_T)$, where P is the pressure, is used to integrate P from simulations to obtain F_{total} with respect to the known ideal-gas free energy. The identity $U_{\text{total}} = ((\partial F_{\text{total}}/T)/(\partial 1/T))_\rho$ is used to integrate, at fixed density, U_{total} from simulations to obtain F_{total} at any desired T . Extrapolations to low T beyond the range of simulations use the form for the potential energies E , $E = E_0 + E_1 T^{\epsilon_2}$, where $E_2 \approx 3/5$ (see refs 8–10). Basin entropies S_{vib} are obtained by averaging over typical inherent structures at each T , whose vibrational frequencies are obtained numerically by diagonalizing the Hessian (matrix of second derivatives of the potential energy). The plot covers at each density the range of T where simulations have been performed. Scaling temperatures T_K^{TI} , the Kauzmann or ideal glass transition temperature where $S_c = 0$, are: 20.44 K, 26.83 K, 34.35 K, 42.49 K and 62.47 K, for $\rho/\rho_0 = 1.1, 1.15, 1.2, 1.25$ and 1.35 respectively. T_K^{PEL} obtained from the PEL method are, in the same order: 20.24 K, 26.1 K, 34.7 K, 44.64 K and 62.35 K.

the basin, contained in the basin free energy $F_{\text{vib}}(\Phi, T)$ and (3) the number density of inherent structures with energy Φ , $\Omega(\Phi)$ (refs 8, 13–15, 18):

$$P(\Phi, T) = \Omega(\Phi) \exp[-\beta(\Phi + F_{\text{vib}}(\Phi, T))]/Q_N(\rho, T) \quad (5)$$

where Q_N , the partition function, normalizes P and is given by $Q_N = \int d\Phi P(\Phi, T)$ over all possible Φ . The configurational entropy density is defined by $S_c(\Phi) \equiv k_B \ln \Omega(\Phi)$ and is related to the T -dependent configurational entropy by $S_c(T) = \int d\Phi S_c(\Phi) P(\Phi, T)$. The quantities $P(\Phi, T)$, F_{vib} and Q_N are obtained from simulations^{8,10,15,18}, and are used to invert the relation in equation (5) to obtain the configurational entropy density $S_c(\Phi)$. The same $S_c(\Phi)$ estimates must result (through not for the same range of Φ) for different T , when the harmonic assumption is valid, which is confirmed by the data shown in Fig. 3a. With increasing density, the maximum $S_c(\Phi)$ value reached decreases, and the distribution becomes broader. It has been argued^{19,30} that the distribution $\Omega(\Phi)$ should be gaussian (equivalently, $S_c(\Phi)$ an inverted parabola). Data in Fig. 3a are fully consistent with this expectation, and are fitted to the form:

$$\frac{S_c(\Phi)}{Nk_B} = \alpha - \frac{(\Phi - \Phi_0)^2}{\sigma^2} \quad (6)$$

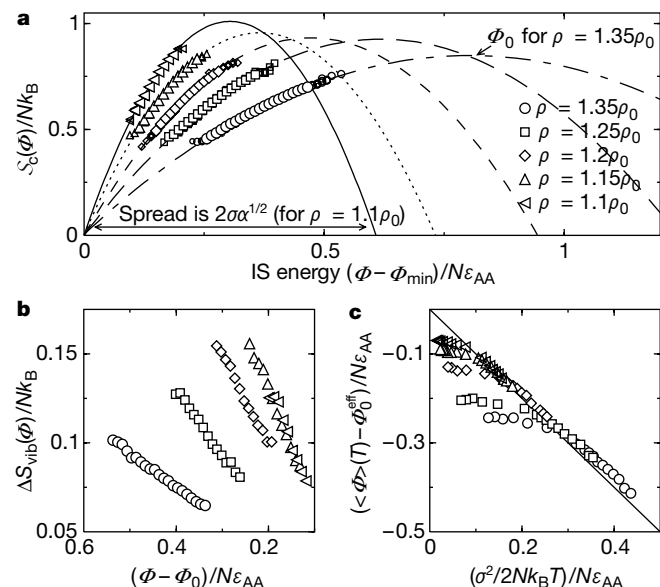


Figure 3 Distribution of energy minima, basin entropy and temperature dependence of inherent structure energies. **a**, The configurational entropy density $S_c(\Phi)$ plotted against inherent structure energy Φ minus the Kauzmann energy $\Phi_{\min}$ below which $S_c(\Phi) = 0$. $\Phi_{\min}$ is the lower root of $S_c(\Phi)/Nk_B = \alpha - (\Phi - \Phi_0)^2/\sigma^2 = 0$. For each density, data from nine temperatures (on average) are shown (smaller symbols) which collapse onto each other, validating the harmonic approximation for calculating basin entropies (see text). Also shown are the averages of these curves (larger symbols), for whose calculation a threshold of 20 samples (out of a total of 1,000) is imposed in each bin. **b**, The T -independent component of the basin entropy $\Delta S_{\text{vib}}(\Phi)$ versus $\Phi - \Phi_0$, where Φ_0 is the value at which $S_c(\Phi)$ is maximum. $\Delta S_{\text{vib}}(\Phi)$ curves are approximately linear. The T independence of $\Delta S_{\text{vib}}(\Phi)$ is an artefact of the classical mechanical treatment used here, but a quantum mechanical calculation of the basin entropy does not alter the picture fundamentally (unpublished results). **c**, Scaled plot of the average IS energies sampled in simulations, which are expected to collapse on the straight line shown at low temperatures (right side of the graph), if the assumption of a gaussian configurational entropy density $S_c(\Phi)$ is valid (see text). The observed data collapse validates the gaussian form used for $S_c(\Phi)$. Deviations at high temperatures are due to the breakdown of harmonicity of the basins. Fitting parameters α , $\Phi_0/N\epsilon_{AA}$, $\sigma/N\epsilon_{AA}$ and $\delta S_{\text{vib}}/k_B$ are: for $\rho = 1.1\rho_0$, 1.010, -6.586, 0.303, -0.721; for $\rho = 1.15\rho_0$, 0.957, -6.684, 0.374, -0.652; for $\rho = 1.2\rho_0$, 0.933, -6.683, 0.489, -0.500; for $\rho = 1.25\rho_0$, 0.927, -6.564, -0.314; and for $\rho = 1.35\rho_0$, 0.849, -6.088, 0.873, -0.188.

where α is the height of the parabola and determines the total number of configurational states, that is, energy minima (the total number is proportional to $\exp(\alpha N)$), Φ_0 and σ^2 respectively define the mean and the variance of the distribution. The fits are displayed in Fig. 3a and the fit parameters listed in the legend. Speedy² has considered a model of the thermodynamics of a glass-former with a gaussian distribution of configurational states, and the basin specific heat that does not depend on the internal parameter (here, the IS energy). He predicts that the fragility of a liquid should be proportional to the logarithm of the number of states, that is to α . From Fig. 3a it is clear that α values, and hence the number of states, decrease with increasing density ρ . But the fragilities (Fig. 1b) show the opposite trend. However, fragility has an additional dependence (implicit in ref. 2) on the variance of the distribution, σ (see equation (6)).

Further, the variation of vibrational frequencies, and hence the basin entropy, from basin to basin, also contributes to the fragility of the liquid. Figure 3b displays the T -independent (see above) basin entropy difference at a given energy Φ with respect to the value at Φ_0 , $\Delta S_{\text{vib}}(\Phi) \equiv S_{\text{vib}}(\Phi, T) - S_{\text{vib}}(\Phi_0, T)$. At all densities, $\Delta S_{\text{vib}}(\Phi)$ displays a clear Φ dependence, which is inconsistent with assumptions in ref. 2. Although such basin dependence of vibrational frequencies and hence the basin entropy have been discussed^{20–29}, with reference to both experimental data and model calculations, its effects on a liquid's fragility have not clearly been established. Figure 3b shows that basins at higher energy (occupied at higher temperature) have lower basin entropies or higher frequencies (see equation (3)). This trend, contrary to what one expects for systems studied experimentally, arises because the constant-density model system here cannot expand when T is raised.

The dependence of $\Delta S_{\text{vib}}(\Phi)$ on Φ is to a great extent linear. Hence, for each density, ΔS_{vib} is fitted to the form $\Delta S_{\text{vib}}(\Phi) = \delta S(\Phi - \Phi_0)$ with the fitting parameters listed in the legend of Fig. 3. The basin free energy, which can be written as $F_{\text{vib}}(\Phi, T) = F_{\text{vib}}(\Phi_0, T) - T\delta S(\Phi - \Phi_0)$, is also linear in Φ . With a gaussian form for Ω , the partition function $Q_N = \int d\Phi P(\Phi, T)$ (P as in equation (5)) is easily evaluated, and results in the following predictions for the T dependence of the average IS energy, and the configurational entropy:

$$\langle\Phi\rangle(T) = \Phi_0^{\text{eff}} - \frac{\sigma^2}{2Nk_B T} \quad (7)$$

where $\Phi_0^{\text{eff}} = \Phi_0 + \sigma^2\delta S/2Nk_B$, and

$$T_S(T) = K_{\text{AG}}^{\text{PEL}}(T)(T/T_K - 1);$$

$$K_{\text{AG}}^{\text{PEL}}(T) = \left(\frac{\sigma\sqrt{\alpha}}{2} + \frac{\sigma^2\delta S}{4Nk_B} \right) \left(1 + \frac{T_K}{T} \right) - \frac{\sigma^2\delta S}{2Nk_B} \quad (8)$$

where $T_K = (\sigma(2Nk_B\sqrt{\alpha} + \sigma\delta S))^{-1}$ is the ideal glass transition temperature where $S_c(T_K) = 0$.

The form above for S_c results in the VFT relation, if $K_{\text{AG}}^{\text{PEL}}(T)$ is constant. This is not quite the case, but the $1/T$ term becomes rapidly irrelevant for $T > T_K$ and $K_{\text{AG}}^{\text{PEL}}(T)$ approaches a constant. Further, T_S values plotted (as points) in Fig. 2b are in very good agreement with the ones from thermodynamic integration, including deviations from linearity for small T/T_K . A check of the overall consistency of the model described here is made in Fig. 3c, which shows a scaled plot of $(\langle\Phi\rangle(T) - \Phi_0^{\text{eff}})$ versus $\sigma^2/2Nk_B T$. From equation (7), one expects $\langle\Phi\rangle(T)$ data for all densities to collapse onto a straight line of (negative) unit slope, with expected deviations at high T due to basin anharmonicity¹⁵. Data in Fig. 3c clearly meet this expectation, although with noticeable deviations for $\rho = 1.35\rho_0$, as also in Fig. 2b.

The expression for $K_{\text{AG}}^{\text{PEL}}$ in equation (8), the thermodynamic fragility index obtained from energy landscape parameters, shows

that fragility depends not only on the multiplicity of states (α) but also on their spread (σ), and how much the basin entropy changes from the lowest to the highest energy basins sampled by the liquid. The extent of this change is quantified by parameters σ and δS (listed in the caption of Fig. 3). Figure 1b shows the average slopes of TS_c (from equation (8) above, in the T range where diffusivities are measured), K_{AG}^{PEL} , which are nearly the same as K_{AG}^{TI} , and in very good agreement with the kinetic fragility index K_{VFT} . This agreement establishes the quantitative relationship between fragility and the energy landscape of the liquid, as K_{AG}^{PEL} is obtained solely in terms of a quantitative description of the liquid's energy landscape. □

Received 31 July; accepted 16 November 2000.

- Angell, C. A. Relaxation in liquids, polymers and plastic crystals—Strong/fragile patterns and problems. *J. Non-Cryst. Solids* **131**–133, 13–31 (1991).
- Speedy, R. J. Relations between a liquid and its glasses. *J. Phys. Chem. B* **103**, 4060–4065 (1999).
- Laughlin, W. T. & Uhlmann, D. R. Viscous flow in simple organic liquids. *J. Phys. Chem.* **76**, 2317–2325 (1972).
- Adam, G. & Gibbs, J. H. On the temperature dependence of cooperative relaxation properties in glass-forming liquids. *J. Chem. Phys.* **43**, 139–146 (1965).
- Angell, C. A. Entropy and fragility in supercooled liquids. *J. Res. NIST* **102**, 171–185 (1997).
- Kob, W. & Andersen, H. C. Testing mode-coupling theory for a supercooled binary Lennard–Jones mixture: The van Hove correlation function. *Phys. Rev. E* **51**, 4626–2641 (1995).
- Sastry, S., Debenedetti, P. G. & Stillinger, F. H. Signatures of distinct dynamical regimes in the energy landscape of a glass-forming liquid. *Nature* **393**, 554–557 (1998).
- Sciortino, F., Kob, W. & Tartaglia, P. Inherent structure entropy of supercooled liquids. *Phys. Rev. Lett.* **83**, 3214–3217 (1999).
- Coluzzi, B., Parisi, G. & Verrocchio, P. Lennard–Jones binary mixture: a thermodynamical approach to glass transition. *J. Chem. Phys.* **112**, 2933–2944 (2000).
- Sastry, S. Liquid limits: The glass transition and liquid–gas spinodal boundaries of metastable liquids. *Phys. Rev. Lett.* **85**, 590–5593 (2000).
- Ito, K., Moynihan, C. T. & Angell, C. A. Thermodynamic determination of fragility in liquids and a fragile-to-strong liquid transition in water. *Nature* **398**, 492–495 (1999).
- Xia, X. & Wolynes, P. G. Fragilities of liquids predicted from the random first order transition theory of glasses. *Proc. Natl Acad. Sci. USA* **97**, 2990–2994 (2000).
- Stillinger, F. H. & Weber, T. A. Packing structures and transitions in liquids and solids. *Science* **225**, 983–989 (1984).
- Stillinger, F. H. A topographic view of supercooled liquids and glass formation. *Science* **267**, 1935–1939 (1995).
- Buechner, S. & Heuer, A. The potential energy landscape of a model glass former: thermodynamics, anharmonicities, and finite size effects. *Phys. Rev. E* **60**, 6507–6518 (1999).
- Speedy, R. J. The hard sphere glass transition. *Mol. Phys.* **95**, 169–178 (1998).
- Scala, A., Starr, F. W., La Nave, E., Sciortino, F. & Stanley, H. E. Configurational entropy and diffusivity of supercooled water. *Nature* **406**, 166–169 (2000).
- Sastry, S. Evaluation of configurational entropy of a model liquid from computer simulations. Proceedings of Unifying Concepts in Glass Physics, Trieste, 1999. *J. Phys. Cond. Mat.* **12**, 6515–6524 (1996).
- Speedy, R. J. & Debenedetti, P. G. The distribution of tetravalent network glasses. *Mol. Phys.* **88**, 1293–1316 (1996).
- Goldstein, M. Viscous liquids and the glass transition. V. Sources of the excess specific heat of the liquid. *J. Chem. Phys.* **64**, 4767–4774 (1976).
- Angell, C. A. & Rao, K. J. Configurational excitations in condensed matter and the bond lattice model for the liquid–glass transition. *J. Chem. Phys.* **57**, 470–481 (1972).
- Angell, C. A. Ten questions on glassformers, and a real space “excitations” model with some answers on fragility and phase transitions. Proceedings of Unifying Concepts in Glass Physics, Trieste, 1999. *J. Phys. Cond. Mat.* **12**, 6463–6476 (2000).
- Green, J. L., Ito, K., Xu, K. & Angell, C. A. Fragility in liquids and polymers: new, simple quantifications and interpretations. *J. Phys. Chem. B* **103**, 3991–3996 (1999).
- Johari, G. P. A resolution for the enigma of a liquid's configurational entropy–molecular kinetics relation. *J. Chem. Phys.* **112**, 8958–8969 (2000).
- Johari, G. P. Contributions to the entropy of a glass and liquid, and the dielectric relaxation time. *J. Chem. Phys.* **112**, 7518–7523 (2000).
- Yamamoto, O. *et al.* Calorimetric study of glassy and liquid toluene and ethylbenzene: thermodynamic approach to spatial heterogeneity in glass-forming molecular liquids. *J. Phys. Chem. B* **102**, 1605–1609 (1998).
- Ngai, K. L. & Yamamoto, O. Thermodynamic fragility and kinetic fragility in supercooled liquids: A missing link in molecular liquids. *J. Chem. Phys.* **111**, 10403–10406 (1999).
- Sciortino, F. & Tartaglia, P. Extension of the fluctuation–dissipation theorem to the physical aging of a model glass-forming liquid. *Phys. Rev. Lett.* (in the press).
- Starr, F. W. *et al.* Thermodynamic and structural aspects of the potential energy surface of simulated water. Preprint <http://arXiv.org/abs/cond-mat/0007487>.
- Heuer, A. & Buechner, S. Why is the density of inherent structures of a Lennard–Jones type system gaussian? Proceedings of Unifying Concepts in Glass Physics, Trieste, 1999. *J. Phys. Cond. Mat.* **12**, 6535–6543 (2000).

Acknowledgements

I thank C. A. Angell, G. P. Johari, K. J. Rao, F. Sciortino, R. Seshadri, R. J. Speedy and U. V. Waghmare for useful discussions and/or comments on the manuscript.

Correspondence should be addressed to the author (e-mail: sastry@jncasr.ac.in).

Observation of shear-induced nematic–isotropic transition in side-chain liquid crystal polymers

Caroline Pujolle-Robic & Laurence Noirez

Laboratoire Léon Brillouin (CEA-CNRS), Ce-Saclay 91191 Gif-sur-Yvette Cédex, France

Flow-induced phase transitions are a fundamental (but poorly understood) property of non-equilibrium systems, and are also of practical importance for tuning the processing conditions for plastics, petroleum products, and other related materials²¹. Recognition that polymers may exhibit liquid crystal properties has led to the discovery of the first tailored side-chain liquid crystal polymers (SCLCPs), which are formed by inserting a spacer between the main polymer chain and the lateral mesogen liquid–crystalline graftings²². Subsequent research has sought to understand the nature of the coupling between the main polymer chain and the mesogens, with a view to obtaining better control of the properties of these tailored structures²². We show here that the parallel or perpendicular orientation of the mesogens with respect to the main chain can be reversed by the application of an external field produced by a shear flow, demonstrating the existence of an isotropic nematic phase transition in SCLCPs. Such a transition, which was theoretically predicted^{1,2} for low-molecular-weight liquid crystals but never observed, is shown to be a general property of SCLCPs too. We expect that these SCLCPs will prove to be good candidate systems for the experimental study of these non-equilibrium phenomena.

Side-chain liquid crystal polymers are made of mesogens (liquid crystal molecules) laterally attached to a conventional main-chain polymer (Fig. 1). The question of the orientation of the main-chain

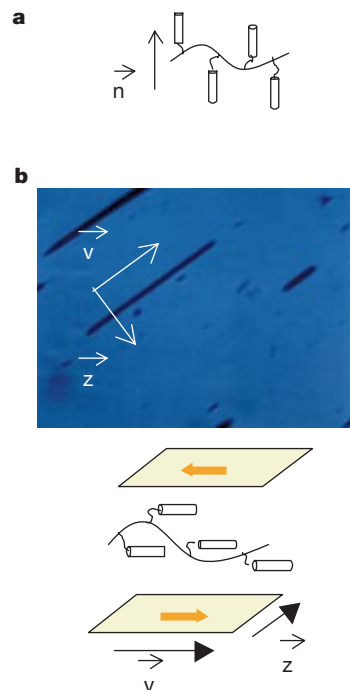
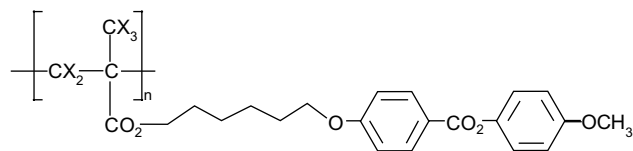


Figure 1 Behaviour of the side-chain liquid crystal polymer described in the text. **a**, Oblate main-chain conformation in the nematic phase, at equilibrium; **b**, strong birefringence due to the emergence of the shear-induced nematic phase (crossed-polarizer micrograph at $\dot{\gamma} = 10 \text{ s}^{-1}$ and $T - T_{NI} = 1.5^\circ\text{C}$), and corresponding shear-induced main-chain conformation. n is the phase direction, v the velocity and z the neutral axis.

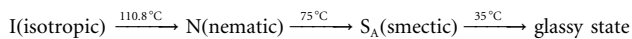
with respect to the mesophase director n is the fundamental issue in SCLCPs. The main-chain/mesogen coupling determines the mesophase type and the conformation of the polymer^{3,4}, and is thus an important parameter in all mechanical and thermodynamic properties. From a theoretical point of view, the orientation of the mesogens with respect to the main-chain results form a set of steric and molecular interactions, which can be treated by a Maier–Saupe approach⁵. In the nematic phase, this interaction potential leads to two fundamental main-chain configurations³ corresponding either to an oblate or a prolate conformation in which the main-chain trajectory is either mostly perpendicular (Fig. 1a) or parallel to the director respectively. Both of these nematic conformations have been experimentally observed⁴. In contrast, in the isotropic phase, no polymer anisotropy can be determined and the scaling laws characterizing this phase exhibit near gaussian or excluded-volume behaviours^{6,7}. Close to the isotropic–nematic (I–N) transition, the phase displays pretransitional fluctuations, with a time-scale about three orders of magnitude larger than in conventional liquid crystals⁸, pointing out the dynamic slowing-down effect due to the mesogen/main-chain coupling.

We show that shearing SCLCPs in the isotropic phase leads to the formation of an aligned mesophase corresponding to a new type of mesogen/main-chain interaction. This transition is the thermotropic equivalent of the well known shear-induced nematic phase of wormlike micellar solutions, extensively studied since its discovery in 1994 (refs 9, 10). More recently, a flow-induced I–N transition for the case of a main-chain thermotropic polyester has been shown by rheo-optical measurements¹¹. However, the main-chain dynamics differ fundamentally from those of the side-chain counterparts. Indeed, the mesogen/main-chain linking in SCLCPs is such that the chain entropy and the mesophase enthalpy can be considered as two independent state variables¹². By studying SCLCPs with different main-chains (acrylate, methacrylate or siloxane) and mesogen structures (cyanobiphenyl or phenylbenzoate substituted with different mesogen extremities), various strengths and signs for the mesogen/main-chain coupling (including oblate⁴) have been tested. In all cases, we have observed this shear-induced transition. Only the critical shear rates vary when the nature of the SCLCP is changed (although they stay in the same order of magnitude) (C.P.-R. and L.N., manuscript in preparation).

The shear-induced transition is illustrated here for one liquid crystal polymethacrylate:



The scaling laws and the linear viscoelasticity characterizing this polymer are presented elsewhere^{6,13}. Two isotopic species (X=H or D) were synthesized and they both display the succession of mesophases:



This liquid crystal polymer, of molecular weight $M_w = 350,000$ and polydispersity index 2.7, is characterized at equilibrium in the nematic phase by a perpendicular mesogen/main-chain coupling⁴ (Fig. 1a).

We first identify the appearance of a shear-induced nematic phase in the isotropic phase using rheo-optical measurements. We define two critical shear rates that limit a two-phase coexistence region. We determine, using small-angle neutron scattering (SANS) under shear, the conformation of the polymer main-chain within the shear-induced phase and show that, in contrast to wormlike micelles systems, the non-equilibrium phase resembles none of

the equilibrium conformations.

Rheo-optical measurements were performed in the (velocity, neutral axis) plane with a custom-built plate–plate shear cell undergoing a backward–forward sawtooth motion. The texture of the nematic phase at rest presents typical Schlieren defects⁵ which disappear above an N–I transition temperature of $T_{NI} = 110.8 \pm 0.05^\circ\text{C}$. No birefringence due to the pretransitional fluctuations could be detected.

In the high-temperature nematic phase ($100^\circ\text{C} < T < T_{NI}$), just below the isotropic phase, the evolution of the texture shows that the application of a shear flow induces macroscopic alignment of the mesophase. The resulting monodomain is uniform and without defects. From the extinction angles of the polarizers, we deduce that the director is aligned along the flow direction. The polymer thus displays a typical flow-aligning behaviour in the high-temperature nematic phase analogous to that displayed by small-molecular-weight liquid crystal molecules.

In the isotropic phase, below a critical shear rate ($\dot{\gamma} < \dot{\gamma}^*$) or away from the I–N transition ($T > T^* > T_{NI}$), no birefringence was observed whatever the angle of the crossed polarizers: the sheared isotropic phase presents the characteristics of an ordinary isotropic phase. Close to T_{NI} , by increasing the shear rate up to a critical value $\dot{\gamma}^*$, a uniform birefringence covered the whole sample. The birefringence was at the maximum when the polarizers were at an extinction angle of $\chi = 45^\circ$; this is a typical flow-aligning behaviour with the director parallel to the shear. The texture was free of defects (Fig. 1b) and optically identical to that displayed by shearing the high-temperature nematic phase; the shear flow has induced a non-equilibrium steady-state nematic phase in the isotropic phase. The birefringence associated with the induced nematic phase appears abruptly as soon as the shear increases up to a critical value $\dot{\gamma}^*$. Conversely, the isotropic phase is immediately restored when the shear is stopped. We verified that the critical value $\dot{\gamma}^*$ is independent of the planar backward–forward sawtooth amplitude by using three different amplitudes (3 mm, 5 mm and 8 mm); the fact that the results can be superposed, together with the abrupt establishment of the birefringence, guarantees that the stationary state is reached.

On increasing the shear rate, a simultaneous increase of the birefringence is observed. Owing to the discontinuous nature of the order parameter at the N–I transition, the apparent increase of the birefringence corresponds to the growth of the induced nematic phase within the “isotropic” phase, or more precisely within a weakly ordered state called a paranematic phase. This phase does not contribute to the birefringence in the (velocity, neutral axis) plane, because the director axis lies in the shear plane at an angle of almost $\pi/4$ with respect to the flow. In addition, the shear-induced biaxiality of the nematic phase is supposed to be negligible in the observation plane. The measured birefringence Δn is thus directly proportional to the nematic phase volume fraction. The growing process continues until the whole sample gap has been converted into an induced nematic phase, which appears as a birefringence saturation plateau. Figure 2a provides an accurate determination of the onset of the birefringence, whereas Fig. 2b, obtained independently, provides the general shape and indicates the saturation at high velocity. The level of the birefringence saturation indicates the temperature dependence of the nematic order parameter which seems to remain constant within $110.8^\circ\text{C} < T < 113^\circ\text{C}$. The values of the birefringence are very high when compared to those for wormlike nematic phases ($\Delta n \approx 10^{-5}$; ref. 14), and they are low compared to those for conventional nematics ($\Delta n \approx 0.3$). They are, however, of the same order as those for nematic elastomers above T_{NI} (ref. 15). Finally, the nematic fraction $\phi_N(\dot{\gamma})$ is quantitatively extracted by dividing the birefringence value at a given shear rate by its value at saturation $\dot{\gamma}_s^*$, that is: $\phi(\dot{\gamma}) = \Delta n(\dot{\gamma})/\Delta n(\dot{\gamma}_s^*)$. The evolution of ϕ_N versus shear rate (equivalent to a renormalization of the previous Δn curves) is remarkably similar to that of the wormlike micelle surfactants¹⁴. The temperature in SCLCPs has an

effect here that is equivalent to that of the inverse of the concentration in the shear-induced transition of wormlike micellar solutions. Another aspect of the existence of the two phases and of the growth of the nematic phase with increasing shear rate, is the presence, in stress measurements, of a plateau considered to be the shear-banding signature. Figure 3 illustrates the stress evolution versus shear rate at a temperature of 117.5 °C for the polymer presented here. This curve shows the existence of a stress plateau observed up to 128 °C. The complete rheological results will be reported elsewhere.

From the rheo-optical measurements, we constructed a temperature–shear rate phase diagram (Fig. 2c). The upper points correspond to the shear at the onset value of the birefringence deduced from Fig. 2a and the lower points to the shear value at saturation deduced from Fig. 2b. The main observation is that the saturation curve does not meet the onset one (higher values cannot be experimentally reached) forbidding any extrapolation to a critical point. We also determined the characteristic time of the induced nematic phase. The transition is induced at a critical shear rate. Thus, the longest characteristic time τ corresponds to the inverse of

the shear rate at the onset value. τ is of the order of a second close to the transition up to 10^{-2} s, eight degrees away from the transition. Compared to the relaxation time of the orientational fluctuations⁸, these values are about 10^3 times larger. Furthermore the experimental results can be modelled by a power-law temperature dependence of the characteristic time: $\tau \propto (T - T_{NI}^*)^{-\nu}$, where the critical exponent is $\nu = 2.3 \pm 0.2$ and $T_{NI}^* - T_{NI} = 2.0 \pm 0.5$ °C (Fig. 2c). This critical exponent deviates from $\nu \cong 1$ expected from the mean-field description and obtained in the case of low-molecular-weight liquid crystal orientational fluctuations. Those results reveal the existence of an additional extra-long-range coupling consistent with the hypothesis of transient network^{16,17}.

We studied the same polymer by SANS to determine whether the polymer main-chain is deformed in the non-equilibrium nematic phase. In the isotropic phase, away from the I–N transition, it was already established that shearing causes no conformational change compared to the equilibrium state: that is, the main chain remains isotropic¹⁸ (at least below 20 s^{-1} , implying that the longest characteristic time is less than 0.05 s). Any interpretation related to a direct coupling with the chain timescale such as in Rouse or reptation dynamics is thus not relevant here²³.

We report here experiments which provide the non-equilibrium conformation of a SCLCP near the I–N transition. Using a specially adapted Couette cell, we determine the main-chain dimensions following the velocity (R_V) and the neutral axis (R_Z) respectively.

The I–N transition is located by the optical change of the texture. The evolution of the main-chain dimensions, one degree above the transition, is shown in Fig. 4b. This figure reveals two shear regimes. At low shear rates ($0 < \dot{\gamma} < 4 \text{ s}^{-1}$), the experimental points indicate that R_V and R_Z remain grouped around the isotropic value. We note that the apparent dispersion of the points is not due to experimental uncertainties (estimated at $\sim 2\%$) but is more probably due to the unstable character of this out-of-equilibrium regime that is also likely at the origin of the irregular macroscopic aspect of the sample. According to Fig. 3c, the shear is supposed to nucleate the nematic phase at $\dot{\gamma}^* \cong 2 \text{ s}^{-1}$. Indeed, as soon as $\dot{\gamma} > 4 \text{ s}^{-1}$, the scattering reveals an alignment of the main chain along the velocity direction (Fig. 4a), the polymer main chain becomes thus oriented parallel to the director (Fig. 1b), and the non-equilibrium conformation corresponds to a parallel main-chain/mesogen coupling. This prolate anisotropy is specific to the shear-induced phase because even very close to the N–I transition, the polymer main chain displays, at rest, the oblate character^{4,6} shown in Fig. 4d. Another indication suggesting a change at the main-chain/mesogen scale was the observation, in the smectic phase, of a decrease of the layer thickness with increasing shear, leading to a tilted smectic phase

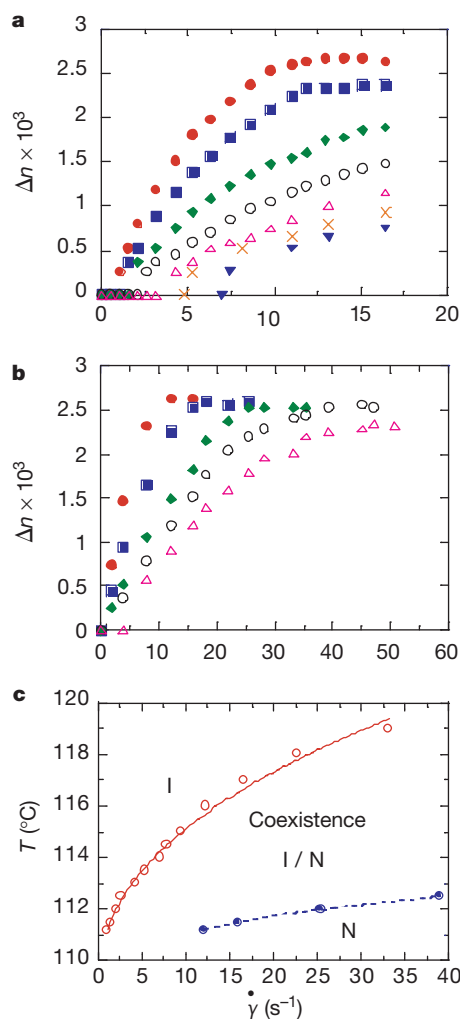


Figure 2 Evolution of the birefringence versus shear rate at different temperatures. Data measured on a sample thickness of $60 \pm 10 \mu\text{m}$ with an intensity accuracy of 6%.

a, Low-shear-rate regime: black circles, 111.2 °C; black squares, 111.5 °C; black diamonds, 112 °C; white circles, 112.5 °C; white triangles, 113 °C; crosses, 113.5 °C; black triangles, 114 °C. **b**, High-shear-rate regime: black circles, 111.2 °C; black squares, 111.5 °C; black diamonds, 112 °C; white circles, 112.5 °C; white triangles, 113 °C. **c**, Non-equilibrium temperature–shear rate phase diagram: white circles, at the onset of the birefringence modelled by a power-law dependence; black circles, at the plateau of the birefringence (the dashed line is a guide to the eye).

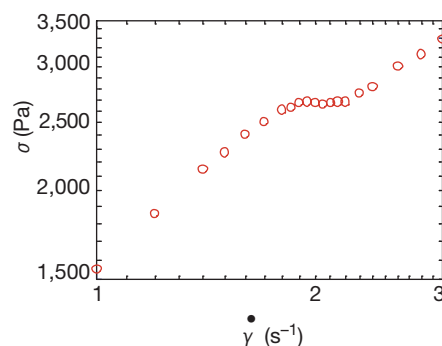


Figure 3 Stress versus shear rate at 117.5 °C measured in a cone-plate cell (diameter: 12 mm, angle: 0.17 rad) in controlled shear rate mode. The critical shear rate (at the onset of the plateau) is obtained for lower shear values and the plateau width is reduced compared to the birefringence results. This discrepancy could originate from the employment of different shear devices and sample thickness. A systematic study of the effect of sample thickness is needed to determine the origin of the difference of the phase diagrams.

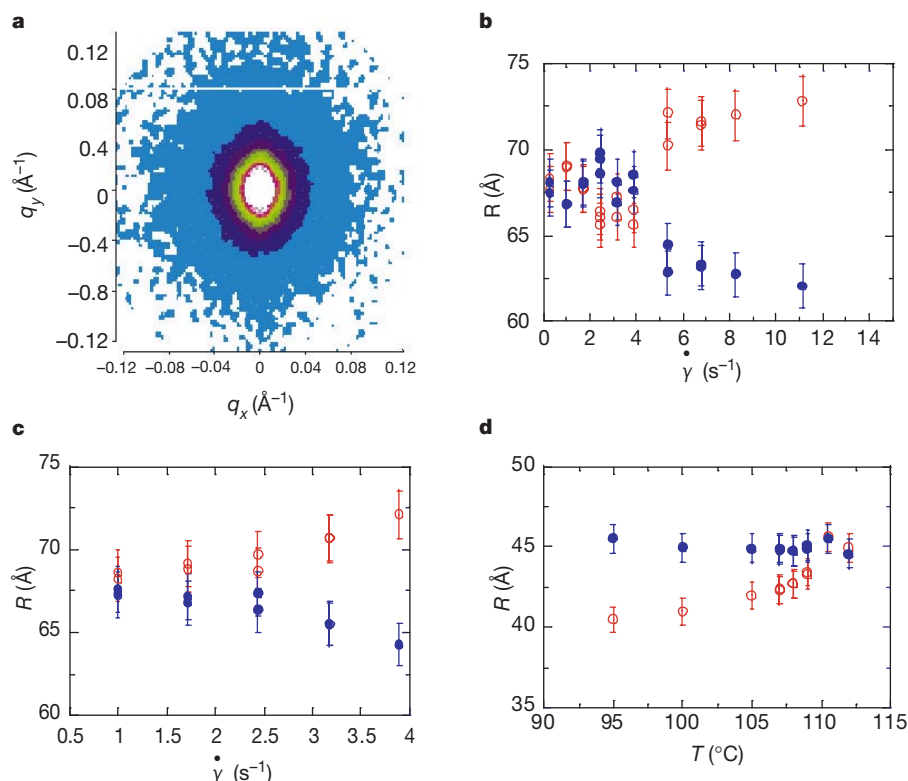


Figure 4 Anisotropy of the main-chain conformation as revealed by neutron scattering. **a**, Scattering pattern displayed by the polymer main-chain in the shear-induced nematic phase at $\dot{\gamma} = 8 \text{ s}^{-1}$ and $T - T_{NI} = 1^\circ\text{C}$. The horizontal axis points in the velocity direction and the vertical axis indicates the neutral axis. The anisotropy of the signal demonstrates that the chain is elongated along the velocity. Polymer main-chain

transition¹⁹ (which differs fundamentally with small liquid-crystal molecules). Figure 4c displays the evolution under shear of the main-chain conformation just below the I–N transition in the ordinary nematic phase. In the absence of external stress (Fig. 4d), the polymer main-chain presents a slightly oblate anisotropy. Under shear flow (Fig. 4c), the reversal of the anisotropy appears clearly; an initially perpendicular coupling is transformed into a non-equilibrium parallel main-chain/mesogen coupling which is conserved across the equilibrium I–N transition.

We have shown that a side-chain polymer can display a shear-induced I–N transition, as displayed by wormlike micelles. We have observed this transition, for similar conditions (shear rates, equivalent birefringence), on different SCLCPs, including both oblate and prolate polymers of various polymerization degrees from 60 to 1,000. The existence of such non-equilibrium states requires strong long-range correlations which are beyond the orientational fluctuation times and polymer dynamics, suggesting the existence of transient networks^{16,17}. At a local scale, the oblate polymer exhibits a transition from perpendicular at equilibrium to parallel main-chain/mesogen orientation under flow. The shear acts on the local interaction potential driving the system to a prolate out-of-equilibrium state against its initial coupling. These results suggest that for strain processes involving SCLCPs or SCLC-networks, such main-chain/mesogen rearrangements, should, at least transiently, occur. □

Methods

We performed the rheo-optical observations in the transmission mode under cross-polarized microscopy (Olympus BX60). The magnification was 10 times and the cross-polarizers were oriented at 45° to the flow direction. The platinum hot-stage (Mettler FP82) ensured a temperature control of $\pm 0.05^\circ\text{C}$. It was equipped with a custom-built parallel plate–plate shear cell executing a planar backward and forward sawtooth motion. The design was based upon the Pieranski–Guyon shear cell²⁰. The displacement amplitude

dimensions: **b**, versus shear rate just above the N–I phase transition ($T - T_{NI} = 1^\circ\text{C}$): white circles, $R_{\parallel}$; black circles, $R_{\perp}$; **c**, versus shear rate below the N–I transition ($T - T_{NI} = -1^\circ\text{C}$): white circles, $R_{\parallel}$; black circles, $R_{\perp}$; and **d**, versus temperature at the approach of the N–I transition at rest ($M_w = 240,000$): white circles, $R_{\parallel}$; black circles, $R_{\perp}$.

can be varied from 3 to 8 mm and the reversal time is estimated as less than 3% of the full displacement time. The transmitted intensity was measured with a photosensitive diode and normalized to the uncrossed polarizer intensity. The sample thickness was constant and estimated at $60 \pm 10 \mu\text{m}$. The birefringence measurements were carried out by selecting one wavelength ($\lambda = 4,500 \text{ \AA} \pm 15 \text{ \AA}$) with a filter. Quantitatively, the birefringence Δn is obtained from the transmitted intensity I/I_0 , by applying the relation $I/I_0 = \sin^2((\pi/\lambda)\delta)$, where λ is the wavelength and δ the retardance, defined by $\delta = \Delta n e$, with e the sample thickness. Using small-angle neutron scattering and specific neutron shear cells, we can access the polymer main-chain deformation in the principal observation planes defining a shear. Our results were extracted from the study of the plane containing the velocity and the neutral axis only. This technique requires the preparation of an isotopic mixture of hydrogenated polymers with polymers deuterated on the main-chain and assumes that the scattering conditions fulfil the Guinier approximation ($qR < 1$)⁴. The signal, $S(q)$, is proportional to the form factor of the main-chain part of the polymer. Let us choose O_x as the axis parallel to the velocity and O_z the axis parallel to the neutral axis. In the $(O_x O_z)$ plane: $S(q) \propto 1/(1 + (q_x^2 R_x^2 + q_z^2 R_z^2))$, where q is the scattering vector ($q = (4\pi/\lambda)\sin(\theta/2)$, with θ the scattering angle). The extracted quantities $R_x = R_{\parallel}$, $R_z = R_{\perp}$ are components of the radius of gyration of the polymer main chain following the two axes. The scattering experiments were carried out on the PAXY spectrometer (Laboratory Léon Brillouin), with a wavelength $\lambda = 8 \text{ \AA}$ and a sample-multidetector distance of 2 m. A specially designed ‘‘Couette’’ cell was constructed in order to ensure the optimized thermal environment required to study pretransitional phenomena; that is, an excellent spatial temperature homogeneity and a negligible temperature gradient. This Couette cell consists of a four-window quartz cylinder. The bulk polymer is placed between the two inner cylinders of diameters 20 mm and 19.2 mm respectively, centred with a concentricity of less than 0.05 mm. The gap is thus 0.4 mm. The internal cylinder is the stator. The external cylinder possesses a double cylindrical window to improve the thermal stability. The heating is provided by a pulsed hot air flow circulating inside the internal cylinder and at the interface between the two external cylinders. The resulting temperature is homogeneous and stable to $\pm 0.02^\circ\text{C}$ over the entire volume of the polymer. The neutron beam trajectory of rectangular section (1 mm width, 7 mm height) was arranged to intercept the rotation axis of the Couette cell, thus allowing the observation of the (velocity, neutral axis) plane. The rotation speed was maintained constant to within 3% using a tachometer.

Received 26 June; accepted 15 November 2000.

1. Olmsted, P. D. & Goldbart, P. Theory of nonequilibrium phase transition for nematic liquid crystals under shear flow. *Phys. Rev. A* **41**, 4578–4581 (1990).

2. Olmsted, P. D. & Goldbart, P. Isotropic-nematic transition in shear flow: state selection, coexistence, phase transitions, and critical behaviour. *Phys. Rev. A* **46**, 4966–4997 (1992).
3. Wang, X. J. & Warner, M. Theory of nematic comb-like polymers. *J. Phys. A* **20**, 713–731 (1987).
4. Noirez, L., Keller, P. & Cotton, J. P. On the structure and the chain conformation of side-chain liquid crystal polymer. *Liq. Cryst.* **18**, 129–148 (1995).
5. de Gennes, P. G. & Prost, J. *The Physics of Liquid Crystals* 2nd edn, 66, 164 (Clarendon, Oxford, 1993).
6. Noirez, L., Daoud-Aladine, A. & Boeffel, C. Scaling laws in side-chain liquid crystalline polymers. *Phys. Rev. Lett.* **80**, 1453–1456 (1998).
7. Fourmeaux-Demange, V., Boué, F., Brûlet, A., Keller, P. & Cotton, J. P. Effect of the molecular weight on the whole conformation of a liquid crystalline comb-like polymer in its melt. *Macromolecules* **31**, 801–806 (1998).
8. Reys, V. *et al.* Short-range-order effects in the isotropic phase of a side-chain polymeric liquid crystal. *Phys. Rev. Lett.* **61**, 2340–2343 (1988).
9. Schmitt, V., Lequeux, F., Pousse, A. & Roux, D. Flow behavior and shear induced transition near an isotropic/nematic transition in equilibrium polymers. *Langmuir* **10**, 955–961 (1994).
10. Berret, J. F., Roux, D. C., Porte, G. & Linder, P. Shear induced isotropic to nematic phase transition in equilibrium polymers. *Europhys. Lett.* **25**, 521–526 (1994).
11. Mather, P. T., Romo-Uribe, A., Han, C. D. & Kim, S. S. Rheo-optical evidence of a flow induced isotropic-nematic transition in a thermotropic liquid crystalline polymer. *Macromolecules* **30**, 7977–7989 (1997).
12. Pleiner, H. & Brand, H. R. Local rotational degrees of freedom in nematic liquid-crystalline side-chain polymers. *Macromolecules* **25**, 895–901 (1992).
13. Rubin, S. F., Kannan, R. M., Kornfield, J. A. & Boeffel, C. Effect of mesophase order and molecular weight on the dynamics of nematic and smectic side-group liquid-crystal polymers. *Macromolecules* **28**, 3521–3530 (1995).
14. Cappellaere, E., Berret, J. F., Decruppe, J. P. & Linder, P. Rheology, birefringence and small-angle neutron scattering in a charged micellar system: evidence of a shear induced phase transition. *Phys. Rev. E* **56**, 1869–1878 (1997).
15. Gleim, W. & Finkelmann, H. Effect of the spacer length on the mechanical coupling between network and nematic order. *Makromol. Chem.* **192**, 2555–2579 (1991).
16. Gallani, J. L., Hilliou, L. & Martinoty, P. Abnormal viscoelastic behavior of side-chain liquid-crystal polymers. *Phys. Rev. Lett.* **72**, 2109–2112 (1994).
17. Brand, H. R. & Kawasaki, K. Mode coupling theory of the isotropic-nematic transition in side-chain liquid crystalline polymers. *J. Phys. II France* **4**, 543–548 (1994).
18. Noirez, L. & Lapp, A. Steady-state shear experiments on a side-chain liquid-crystal polymer: determination of the polymer conformation and liquid-crystal structure. *Phys. Rev. E* **53**, 6115–6120 (1996).
19. Noirez, L. Shear-induced S_A - S_C transition in side-chain liquid-crystalline polymers. *Phys. Rev. Lett.* **84**, 2164–2167 (2000).
20. Pieranski, P. & Guyon, E. Instability of certain flows in nematic liquids. *Phys. Rev. A* **9**, 404–417 (1974).
21. Dealy, J. H. & Wisbrun, K. *Melt Rheology and its Role in Plastics Processing* (Van Nostrand Reinhold, New York, 1990).
22. Finkelmann, H., Ringsdorf, H., Sjol, W. & Wendorff, J. H. Model consideration and examples of enantiotropic liquid crystalline polymers. *Makromol. Chem.* **179**, 273 (1978).
23. de Gemmes, P. G. *Scaling Concepts in Polymer Physics* 165, 224 (Cornell Univ. Press, 1979).

Acknowledgements

We thank P. Baroni for the construction of the shear devices and his technical help during the experiments.

Correspondence and requests for materials should be addressed to L.N.
(e-mail: noirez@llb.saclay.cea.fr).

An abrupt climate event in a coupled ocean–atmosphere simulation without external forcing

Alex Hall* & Ronald J. Stouffer†

* Lamont-Doherty Earth Observatory, Palisades, New York 10964, USA

† Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey 08542, USA

Temperature reconstructions from the North Atlantic region indicate frequent abrupt and severe climate fluctuations during the last glacial and Holocene periods^{1–3}. The driving forces for these events are unclear and coupled atmosphere–ocean models of global circulation have only simulated such events by inserting large amounts of fresh water into the northern North Atlantic Ocean^{4,5}. Here we report a drastic cooling event in a 15,000-yr

simulation of global circulation with present-day climate conditions without the use of such external forcing. In our simulation, the annual average surface temperature near southern Greenland spontaneously fell 6–10 standard deviations below its mean value for a period of 30–40 yr. The event was triggered by a persistent northwesterly wind that transported large amounts of buoyant cold and fresh water into the northern North Atlantic Ocean. Oceanic convection shut down in response to this flow, concentrating the entire cooling of the northern North Atlantic by the colder atmosphere in the uppermost ocean layer. Given the similarity between our simulation and observed records of rapid cooling events, our results indicate that internal atmospheric variability alone could have generated the extreme climate disruptions in this region.

An extreme event is evident in the time series of surface air temperature (SAT) over southeastern Greenland and the neighbouring oceans (Figs 1a and c). Occurring around model year 3100, it lasts approximately 40 yr and is characterized by cooling of about 3 °C. As it is more than six standard deviations (σ) from the mean at its peak, the cooling can objectively be considered extreme. If the time series of annual mean SAT values in this region were normally distributed, such an anomaly would occur once every few hundred million years. Also, the probability that such a large cooling would last so long is extremely small. If the variations that last for less than 40 yr are removed from the time series, the SAT at the event's peak lies 10–11 σ below the mean.

The source of this extraordinary lower-atmospheric cooling is a precipitous drop in underlying sea surface temperature (SST). Beginning around model year 3086, a modest cold SST anomaly of about 0.5 °C develops off central Greenland's east coast. This cooling intensifies until model year 3100, reaching 2.0–2.5 °C. Then it expands southwest along Greenland's entire east coast (Fig. 2a), reaching a peak magnitude of about 4.0 °C. The annual mean SST anomalies during this phase are 8–10 σ below the mean. The evolution of sea surface salinity (SSS) anomalies is similar. At the event's peak (Fig. 2b), large amounts of fresh water appear along Greenland's entire southeastern coast. This SSS anomaly also lies 8–10 σ below the mean.

These surface anomalies can be traced to an abnormally intense East Greenland current. Flowing southward along Greenland's east coast in the model as in reality⁶, this current transports relatively cold and fresh Arctic water into the northern North Atlantic Ocean. At around model year 3083, the current gradually intensifies off central Greenland, leading to cooling and freshening there. Then, around model year 3096, the intensification expands, reaching Greenland's tip by year 3100 (Fig. 2d). This leads to the massive cooling and freshening along Greenland's southeast coast at the event's peak. By model year 3120, the current anomaly disappears.

The intensity of the East Greenland current at this stage is unusual, partly explaining the unusual drops in SST and SSS. Measured through the southward component of the surface current across the Denmark Strait (66.7° N, 13.1 to 28.1° W), the annual mean current intensity rises about 4.5 σ above the mean. When variations of duration less than 40 yr are removed, the intensity is about 6 σ above the mean, indicating that the current's strength is even more unusual when its long duration is taken into account.

Cooling and freshening off central Greenland's east coast and the intensification of the East Greenland current are characteristics of the model's 'great salinity anomalies' (GSAs)—a quasi-periodic 40–80-yr oscillation already documented in this model⁷. A typical simulated GSA involves a cooling of about 1–1.5 °C and freshening of about 0.2–0.3 practical salinity units, psu, off the central Greenland coast. This is broadly consistent with actual GSAs that occurred in the late 1960s and 1970s, and possibly also in the early part of the twentieth century⁸. Early in its lifetime, the extreme event resembles a typical simulated GSA. However, by model year 3101, it differs significantly in magnitude and in geographical extent (Figs 2a and

b). Whereas typical simulated GSAs are centred between Iceland and Greenland, the extreme event's SST and SSS anomalies extend from there to beyond the tip of Greenland.

Why is this GSA so extreme? The explanation lies partly in the reason for the unusually intense East Greenland current. Surface winds strengthen this current, transforming a typical GSA into the extreme event. The surface pressure anomaly pattern during the extreme event (Fig. 3) drives northwesterly winds along Greenland's east coast. These winds, in turn, induce a southwestward Ekman current in the surface ocean. In fact, the strongest northwesterly winds over Greenland's coast in the 15,000-yr experiment coincide exactly with the extreme event. (Figure 4 shows information on northwesterly wind index definition and data processing.) The winds begin intensifying around model year 3080, reaching values about $4\text{--}5\sigma$ above the mean at around year 3100. By year 3120, they subside. This strongly indicates that the unusually intense northwesterly winds are the trigger for the extreme event.

Other unusually large but less extreme GSAs occur in this experiment and are also linked to abnormally strong northwesterly winds. The SST variability distribution in the Denmark Strait (Fig. 4) reveals several instances of improbably cold SST but not correspondingly large warm events. Whereas no warm anomaly exceeds 3σ , the SST falls more than 3σ below the mean for a total of 163 yr, of which 33 yr are attributable to the extreme event. These represent a total of 10 separate events where the SST is unusually cold for several years at a time. During all these years, the northwesterly winds blowing across the Denmark Strait are stronger than normal. There is a less robust correspondence between more typical GSAs and wind strength. The correlation between the entire SST and wind strength time series is -0.55 , but the correlation falls to -0.39 if times when SST fluctuations are greater than 1σ from the mean are removed from the two time series. Wind variations certainly generate highly unusual GSAs through their effect on surface currents, but they probably play a weaker role in more typical GSAs.

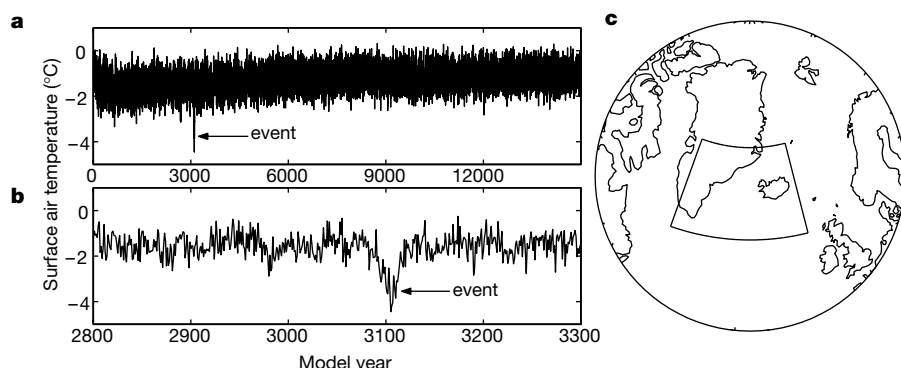


Figure 1 Time series of surface air temperature averaged over southeast Greenland and the neighbouring ocean ($57.8\text{--}71.1^\circ\text{N}$, $11.3\text{--}48.8^\circ\text{W}$). **a**, Entire model integration;

b, model years 2800–3300 only. **c**, The box encloses the region used for averaging in **a** and **b**.

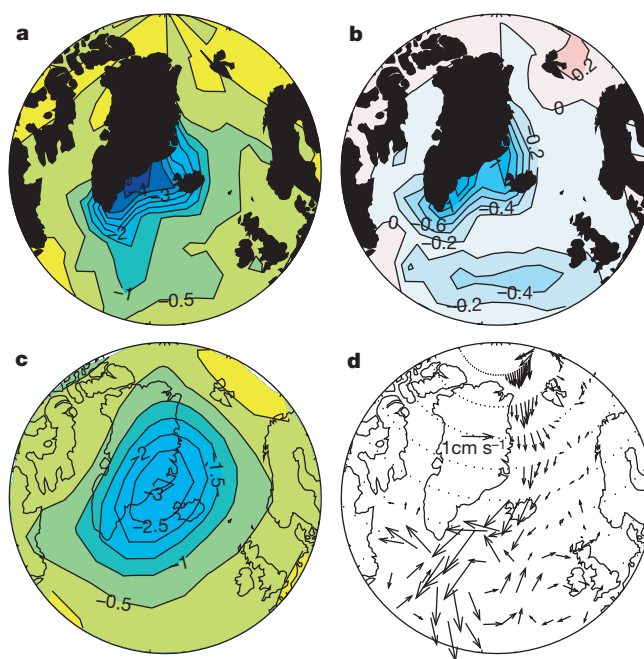


Figure 2 Sea surface temperature and salinity, surface air temperature and ocean current anomalies. **a**, SST (°C); **b**, SSS (practical salinity units or psu); **c**, SAT (°C); and **d**, surface ocean currents averaged over model years 3101–3110, the approximate decade when the event reaches its peak. All anomalies are calculated relative to the first 5,000 yr of the experiment. Real rather than model geography is used to facilitate orientation. A few

hundred kilometres wide, the model's East Greenland current is qualitatively realistic, transporting on average about 4 sverdrups of Arctic water southward along Greenland's coast. However, due to the model's coarse resolution, the simulated current is wider and slower than the observed current, which transports about 8 sverdrups of water southward in a narrow coastal strip about 150 km wide⁶.

SST in this region is very sensitive to northwesterly wind fluctuations when the winds are stronger than normal (Fig. 4). This is due to a climate instability unique to this region. The cold and fresh surface waters entering the area off Greenland's coast during unusual GSAs are very buoyant. During the extreme event, annual mean surface density, a function of temperature and salinity, falls 8–10 σ below the mean in the area where SST and SSS are unusually low. Thus, the density-reducing effects of low SSS overwhelm the density-increasing effects of cold SST. Because of their extreme buoyancy, the cold, fresh waters eliminate deep convection off Greenland's southeastern coast. Like other high-latitude locations, the frigid atmosphere in this region constantly absorbs heat from the relatively warm ocean. The cutoff of surface waters from underlying layers during a convection shutdown sharply reduces the ocean's effective heat capacity. This concentrates the entire atmospheric cooling of the ocean in the first layer, amplifying the cold SST of invading northern waters. This positive feedback does not operate when northwesterly winds are unusually feeble, the East Greenland current weakens, and SST rises above normal (Fig. 4). Because SSS and hence surface density are also high at these times, convection may be enhanced; this does little, however, to alter the effective heat capacity of the ocean, because deep convection already exists here. This asymmetry explains why SST is affected to a much greater extent by wind-driven intensification rather than by weakening of the East Greenland current. This positive feedback would also operate if these waters became fresher because of a large meltwater discharge from ice sheets. In fact, much of this region's initial cooling in numerical experiments simulating such events⁵ may be due to a convection shutdown.

This asymmetrical positive feedback explains how a merely unusual wind intensification produces a much more extreme variation in SST. The probability of finding a wind anomaly 4–5 σ below the mean in a 15,000-yr normally distributed time series is not negligible. For example, we would expect an anomaly 4 σ from

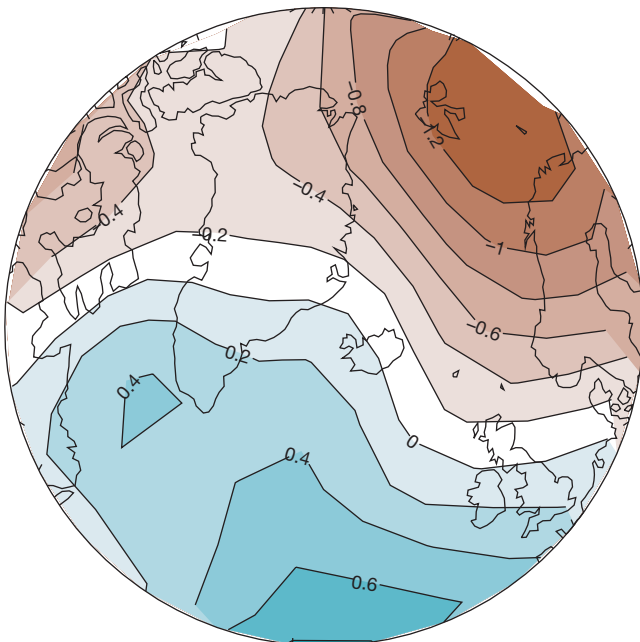


Figure 3 The surface pressure anomalies (in hPa) relative to the first 5,000 yr of the experiment averaged over model years 3096–3110. This pattern is typical of the surface pressure anomalies during most of the extreme event. As the simulated climatological winds over Greenland blow predominantly from the west, the winds driven by this pressure pattern represent both an increase in wind strength and a slight change in direction towards the south.

the mean to occur approximately once every 7,500 yr. Moreover, the northwesterly wind index time series contains equal amounts of variability on all timescales longer than a few days. This suggests that wind variability in this region is randomly generated by the model atmosphere, rather than induced in some way by local SST variability, which has a spectral peak at the 40–80-yr timescale. In addition, the low-frequency component of the model's North Atlantic Oscillation (NAO), an internal mode of the atmosphere, is positive during the entire extreme event (Fig. 4 has details on the NAO index). The NAO's positive phase is associated with abnormally strong westerly winds over the North Atlantic from 45–70° N. The unusual intensity of the northwesterly winds blowing across Greenland during the event thus stems partly from the positive NAO. For these reasons, the wind fluctuation that triggers the extreme event probably results from the model atmosphere's internal variability.

This unusual wind fluctuation affects the climate elsewhere by moving warm maritime air over the Eurasian continent. This causes a widespread SAT increase of about 0.5 °C over Europe and

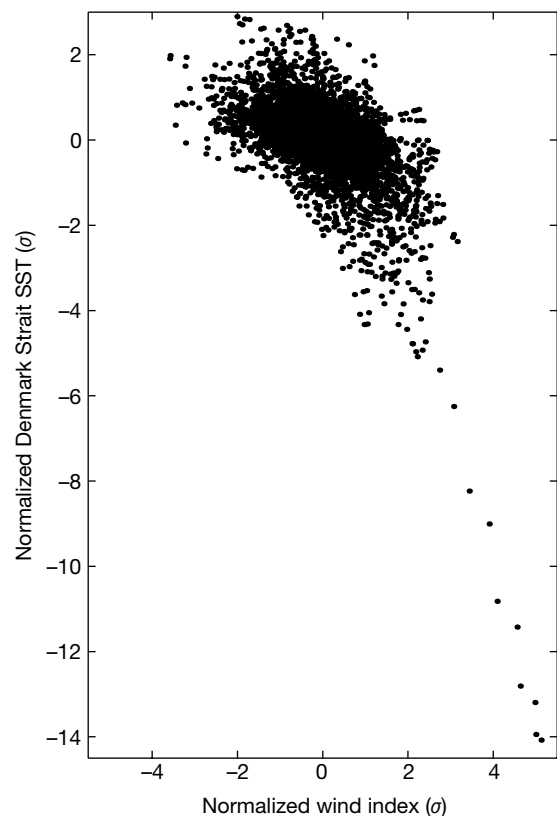


Figure 4 Scatter plot of sea surface temperature (SST) averaged over the Denmark Strait (62.2–66.7° N, 18.8–30° W) versus winds blowing perpendicular to the Denmark Strait. Before plotting, SST was made to lag the wind index by 2 yr. The ten points representing the coldest SSTs all correspond to the extreme event. The wind index is defined as the surface pressure difference between two regions: 48.9–57.8° N, 33.8–56.25° W (an area south of the tip of Greenland) and 66.7–75.6° N, 11.3–41.3° E (centred over northern Scandinavia). A Hamming 40-yr low-pass filter has been applied to the SST and wind index, and both time series have been normalized by their respective standard deviations. For clarity, only every third year of the time series is shown. The model climate exhibits a quasi-linear millennial-scale trend over the course of the 15,000-yr experiment (Fig. 1a). To ensure this model drift does not distort the distribution of variability presented here, all variations lasting longer than 1,000 yr were removed from the SST and wind index time series using a Hamming high-pass filter. The North Atlantic Oscillation index is defined as the pressure difference between two regions: 40.0–48.9° N, 11.2–26.2° W and 66.7–75.6° N, 3.8–18.8° W. Variations lasting less than 40 yr were also removed from this time series.

central Eurasia at the event's peak (model years 3101–3110). These temperatures are the warmest of the entire low-frequency (timescale ≥ 40 yr) time series of SAT averaged over this region, rising about 3.5σ above the mean. Although it is less extreme than the simultaneous cooling over southern Greenland (Fig. 2c), this warming covers a much wider area, so that the extreme event's signature is not visible in a hemispheric or global mean time series of SAT.

The simulated extreme event is qualitatively similar to some abrupt millennial-scale North Atlantic cooling events seen in the Holocene record¹. The observed episodes, for which there is no known external forcing, are characterized by intensification of the East Greenland current, transport of water from the north of Iceland into the northern North Atlantic, and drops in SST of about 2°C . In addition, the observed millennial-scale events have a signal in the low-latitude North Atlantic⁹. The simulated event also has a signal throughout the North Atlantic: the buoyant waters invading the northern North Atlantic slow the model's North Atlantic overturning, beginning around model year 3104. By year 3115, the thermohaline circulation weakens by about 16% from its mean value of 17.6 sverdrups (a 3.5σ anomaly). This reduces SST throughout the North Atlantic by a few tenths of a degree. Because the thermohaline circulation weakening occurs well after the event's onset, it is a response to the event rather than its cause.

The real millennial-scale events differ from the simulated event in that they probably last somewhat longer than 30–40 yr, although their exact length is uncertain since the marine sediments supplying information about them are subject to bioturbation. They also recur roughly every 1,500 yr (ref. 1). The simulated extreme event only occurred once during the entire experiment, so similar events would probably be rarer than the observed events if the experiment were continued for longer than 15,000 yr. These differences could be attributable to a deficiency in the model—the coarse resolution of the ocean model forces us to use unrealistically high horizontal and vertical diffusivity^{10,11}. This probably makes the North Atlantic climate more stable than it should be. The massive outflow of cold and fresh water during the extreme event diffuses too quickly into the surrounding warm salty water, shortening the event's lifetime and limiting its geographical extent. The other simulated cold events where Denmark Strait SST falls more than 3σ below the mean might have developed further, lasted longer, and covered a larger area if the model ocean were less diffusive. As we noted above, over the course of the 15,000-yr experiment there are ten such events.

Although the probability of this type of drastic event happening in the near future seems low, it may be increasing for two reasons. First, nearly all coupled ocean–atmosphere models show an increase in atmospheric fresh water transport to the Arctic of about a tenth of a sverdrup when atmospheric CO_2 is doubled¹². This significantly freshens the Arctic. Because the mechanism for the extreme event relies on fresh water transport out of the Arctic, increasing greenhouse gas concentrations could make a similar event more likely. Second, consistent with the observed 20–30-yr increase in the real climate's NAO index is a strengthening of the northwesterly winds blowing perpendicular to the East Greenland current^{13,14}. If these winds continue to strengthen, they could produce a drastic climate event in the northern North Atlantic. □

Methods

The coupled ocean–atmosphere model¹⁵ has a global domain with realistic geography. It has been used to study global warming^{15–18}, climate variability^{19,20} and palaeoclimate^{4,5,21}. The atmospheric component has nine vertical levels, with horizontal distributions of variables represented by spherical harmonics and grid-point values²² with 4.5° latitude by 7.5° longitude spacing. Sunshine varies seasonally, but not daily. Cloud cover is predicted on the basis of relative humidity. A land surface model computes surface fluxes of heat and water²³. The ocean component²⁴ employs a finite-difference technique using a grid system with approximately 4.5° latitude by 3.8° longitude spacing. There are 12 vertical levels in the ocean. The effects of mesoscale eddies are represented by diffusion of potential

temperature and salinity on constant-density surfaces. A sea-ice model computes ice thickness on the basis of thermodynamic heat balance and movement of ice by ocean currents. The atmospheric and oceanic components exchange heat, water, and momentum once a day. The 15,000-yr experiment is designed to be a perpetual simulation of the present-day climate. Accordingly, carbon dioxide, the only greenhouse gas other than water vapour in the model, is fixed to present-day levels, and solar radiation does not vary from year to year. Continental ice sheets are fixed to their present-day positions. Glacial dynamics therefore play no role in the extreme event or any variability simulated by this model.

The coupled model's quasi-equilibrium initial condition is obtained by separate integrations of the atmospheric and oceanic components using observed surface boundary conditions (SST, SSS and sea-ice thickness). When the integration of the coupled model begins from this initial condition, the simulated climate drifts toward a less realistic equilibrium state. To reduce drift, the fluxes of heat and water at the ocean–atmosphere interface are adjusted by given amounts before they are imposed on the ocean surface. The flux adjustments are determined before the coupled model run from separate integrations of the atmosphere and ocean models. They do not change from one year to the next and are independent of the SST and SSS anomalies that develop during the simulation. Thus they neither damp nor amplify these anomalies in any systematic way. Although the adjustments do not eliminate the shortcomings of the model²⁵, they do help to prevent rapid drift from the initial state. Such drift could seriously distort the internal variability that is the subject of this study.

Received 4 September; accepted 24 November 2000.

- Bond, G. et al. A pervasive millennial-scale cycle in North Atlantic Holocene and glacial climates. *Science* **278**, 1257–1266 (1997).
- Dansgaard, W. et al. A new Greenland deep ice core. *Science* **218**, 1273–1277 (1982).
- Johnsen, S. J. et al. Irregular glacial interstadials recorded in a new Greenland ice core. *Nature* **359**, 311–313 (1992).
- Manabe, S. & Stouffer, R. J. Simulation of abrupt climate change induced by freshwater input to the North Atlantic Ocean. *Nature* **378**, 165–167 (1995).
- Manabe, S. & Stouffer, R. J. Coupled ocean–atmosphere model response to fresh-water input: Comparison to Younger Dryas event. *Paleoceanography* **12**, 321–336 (1997).
- Woodgate, R. A., Fahrbach, E. & Rohardt, G. Structure and transports of the East Greenland Current at 75N from moored current meters. *J. Geophys. Res.* **104**, 18059–18072 (1999).
- Delworth, T., Manabe, S. & Stouffer, R. J. Multidecadal climate variability in the Greenland Sea and surrounding regions: a coupled model simulation. *Geophys. Res. Lett.* **24**, 257–260 (1997).
- Dickson, R. R., Meinke, J., Malmberg, S. A. & Lee, A. J. The “Great Salinity Anomaly” in the northern North Atlantic 1968–1982. *Prog. Oceanogr.* **20**, 103–151 (1988).
- deMenocal, P., Ortiz, J., Guilderson, T. & Sarnthein, M. Coherent high and low latitude climate variability during the Holocene warm period. *Science* **288**, 2198–2202 (2000).
- England, M. H. Using chlorofluorocarbons to assess ocean climate models. *Geophys. Res. Lett.* **22**, 3051–3054 (1995).
- Rabaille, D. Y. & Weaver, A. J. Validation of sub-grid-scale mixing schemes using CFCs in a global model. *Geophys. Res. Lett.* **22**, 2917–2920 (1995).
- Kattenberg, A. et al. in *Climate Change 1995: The Science of Climate Change* 285–357 (Cambridge Univ. Press, Cambridge, 1996).
- Hurrell, J. W. Decadal trends in the North Atlantic Oscillation regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
- Thompson, D. W. J., Wallace, J. M. & Hegerl, G. C. Annular modes in the extratropical circulation. Part II: trends. *J. Clim.* **13**, 1018–1036 (2000).
- Manabe, S., Stouffer, R. J., Spelman, M. & Bryan, K. Transient responses of a coupled ocean–atmosphere model to gradual changes of atmospheric CO_2 . Part I: annual-mean response. *J. Clim.* **4**, 785–817 (1991).
- Manabe, S., Spelman, M. J. & Stouffer, R. J. Transient response of a coupled ocean–atmosphere model to gradual changes of atmospheric CO_2 . Part II: Seasonal response. *J. Clim.* **5**, 105–126 (1992).
- Manabe, S. & Stouffer, R. J. Multiple-century response of a coupled ocean–atmosphere model to an increase of atmospheric carbon dioxide. *J. Clim.* **7**, 5–23 (1994).
- Stouffer, R. J. & Manabe, S. Response of a coupled ocean–atmosphere model to increasing atmospheric carbon dioxide: Sensitivity to the rate of increase. *J. Clim.* **12**, 2224–2237 (1999).
- Delworth, T., Manabe, S. & Stouffer, R. J. Interdecadal variation of the thermohaline circulation in a coupled ocean–atmosphere model. *J. Clim.* **6**, 1993–2011 (1993).
- Manabe, S. & Stouffer, R. J. Low frequency variability of surface air temperature in a 1000-year integration of a coupled ocean–atmosphere model. *J. Clim.* **9**, 376–393 (1996).
- Manabe, S. & Stouffer, R. J. Two stable equilibria of a coupled ocean–atmosphere model. *J. Clim.* **1**, 841–866 (1988).
- Gordon, C. T. & Stern, W. A description of the GFDL Global Spectral Model. *Mon. Weath. Rev.* **110**, 625–644 (1982).
- Manabe, S. Climate and the ocean circulation: I. The atmospheric circulation and the hydrology of the Earth's surface. *Mon. Weath. Rev.* **97**, 739–774 (1969).
- Bryan, K. & Lewis, L. A water mass model of the world ocean. *J. Geophys. Res.* **84**, 2503–2517 (1979).
- Marotzke, J. & Stone, P. Atmospheric transports, the thermohaline circulation, and flux adjustments in a simple coupled model. *J. Phys. Oceanogr.* **25**, 1350–1364 (1995).

Acknowledgements

A.H. is supported by a Lamont Postdoctoral Fellowship. We thank G. Bond for useful discussions, and T. Delworth, P. deMenocal, I. Held, M. Elliot, M. Latif, J. Mahlman, S. Manabe and R. Wood for comments on the manuscript.

Correspondence and requests for materials should be addressed to A.H. (e-mail: alexhall@ldeo.columbia.edu).

Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago

Simon A. Wilde*, John W. Valley†, William H. Peck‡§ & Colin M. Graham§

* School of Applied Geology, Curtin University of Technology, GPO Box U1987, Perth, Australia

† Department of Geology & Geophysics, University of Wisconsin, Madison, Wisconsin 53706, USA

§ Department of Geology & Geophysics, University of Edinburgh, Edinburgh, EH9 3JW, UK

‡ Present address: Department of Geology, Colgate University, Hamilton, New York 13346, USA

No crustal rocks are known to have survived since the time of the intense meteor bombardment that affected Earth¹ between its formation about 4,550 Myr ago and 4,030 Myr, the age of the oldest known components in the Acasta Gneiss of northwestern Canada². But evidence of an even older crust is provided by detrital zircons in metamorphosed sediments at Mt Narryer³ and Jack Hills^{4–8} in the Narryer Gneiss Terrane⁹, Yilgarn Craton, Western Australia, where grains as old as ~4,276 Myr have been found⁴. Here we report, based on a detailed micro-analytical study of Jack Hills zircons¹⁰, the discovery of a detrital zircon with an age as old as $4,404 \pm 8$ Myr—about 130 million years older than any previously identified on Earth. We found that the zircon is zoned with respect to rare earth elements and oxygen isotope ratios ($\delta^{18}\text{O}$ values from 7.4 to 5.0‰), indicating that it formed from an evolving magmatic source. The evolved chemistry, high $\delta^{18}\text{O}$ value and micro-inclusions of SiO_2 are consistent with growth from a granitic melt^{11,2} with a $\delta^{18}\text{O}$ value from 8.5 to 9.5‰. Magmatic oxygen isotope ratios in this range point toward the involvement of supracrustal material that has undergone low-temperature interaction with a liquid hydrosphere. This zircon thus represents the earliest evidence for continental crust and oceans on the Earth.

We extracted a new aliquot of zircon from the original heavy mineral concentrate of the Jack Hills conglomerate from which the oldest documented crystal (~4,276 Myr; ref. 4) was obtained, and identified several grains with ages in excess of 4 Gyr. One of these is a deep purple zircon (W74/2-36), measuring 220 by 160 μm , that is a broken fragment of a larger crystal (Fig. 1). It shows few internal

complexities or inclusions, a feature previously noted in zircons older than 4 Gyr from the Jack Hills conglomerate^{4,13}.

We performed U-Th-Pb zircon analyses on the Perth Consortium SHRIMP II ion microprobe at Curtin University of Technology in Western Australia, following standard operating techniques^{14–16}. Six analyses of grain W74/2-36 were initially made and five of these gave $^{207}\text{Pb}/^{206}\text{Pb}$ ages in excess of 4.3 Gyr (Table 1); the oldest and most concordant (97% concordance, Table 1) had an age of $4,363 \pm 8$ Myr (2σ), approximately 90 Myr older than the oldest known terrestrial material⁴. Other sites yielded more discordant ages and fell on a discordia line that passed through zero, indicative of recent lead loss (Fig. 2). A detailed cathodoluminescence (CL) image made after analysis indicated that all but the oldest site overlapped cracks within the crystal (Fig. 1a). However, crack-free domains greater than 50 μm in diameter were present and, after oxygen isotope and rare earth element (REE) analysis, approximately 20 μm was ground off the surface of the zircon to remove all previous analytical pits; the grain was then re-analysed on SHRIMP II in an attempt to obtain more concordant data. Care was taken to locate new analytical sites away from identifiable cracks and this proved partially successful, with four analyses obtained from such areas (Fig. 1b). Three of these sites yielded data that are 100, 98 and 95% concordant (Table 1), and confirm the initial estimate of the age, giving $^{207}\text{Pb}/^{206}\text{Pb}$ ages of $4,355 \pm 4$ Myr, $4,341 \pm 6$ Myr and $4,364 \pm 6$ Myr (2σ), respectively (Table 1 and Fig. 2). Importantly, the fourth site, located near the pointed, broken termination of the crystal (right corner of Fig. 1b), was also nearly concordant (99%) and this gave a $^{207}\text{Pb}/^{206}\text{Pb}$ age of $4,404 \pm 8$ Myr (2σ) (Fig. 2), about 40 Myr older than the other results and only about 150 Myr younger than the oldest high-temperature inclusions known in meteorites¹⁷ (~4,560 Myr), which constrain the maximum age for the Earth.

The chemistry of this crystal therefore provides important insights into early crustal processes. Because the internal cracks that are evident in CL and back-scattered electron (BSE) images (Fig. 1) may affect the chemistry, they were investigated using a high-resolution scanning electron microscope (SEM) and electron microprobe. Unlike the cracks in the xenocrysts older than 4 Gyr that were recently discovered in granitoids from the Narryer and Murchison Terranes of Western Australia¹⁸, which are infilled with recrystallized zircon, the cracks in W74/2-36 are open. They contain thin (generally less than 300 nm thick), discontinuous films of phosphates, resulting in slightly elevated values of Fe, Al, P and Y. These are, however, too small to have a significant effect on the oxygen budget of 25- μm analytical sites; thus they do not compromise our interpretation of the oxygen data.

Grain W74/2-36 has U and Th values ranging from 127–487 and

Table 1 Shrimp U-Pb-Th isotopic analytical data

Spot	U (p.p.m.)	Th (p.p.m.)	Th/U ratio	Pb (p.p.m.)	f206%	²⁰⁴ Pb/ ²⁰⁶ Pb	²⁰⁷ Pb*/ ²⁰⁶ Pb*	²⁰⁸ Pb*/ ²³² Th	²⁰⁸ Pb*/ ²⁰⁶ Pb*	²⁰⁶ Pb*/ ²³⁸ U	²⁰⁷ Pb*/ ²³⁵ U	Conc. (%)	²⁰⁶ Pb*/ ²³⁸ U age (Myr)	²⁰⁷ Pb*/ ²³⁵ U age (Myr)	²⁰⁸ Pb*/ ²³² Th age (Myr)	²⁰⁷ Pb*/ ²⁰⁶ Pb age (Myr)
Session 1																
74-36	258	178	0.69	355	0.22	0.00014	0.5432 ± 1	0.222 ± 5	0.1645 ± 12	0.928 ± 18	69.5 ± 1.4	97	4,233 ± 61	4,321 ± 20	4,058 ± 80	4,363 ± 4
74-37	449	347	0.77	615	0.048	0.00003	0.5386 ± 1	0.225 ± 5	0.1895 ± 8	0.919 ± 18	68.2 ± 1.3	97	4,201 ± 59	4,303 ± 20	4,109 ± 75	4,350 ± 3
74-38	450	1,224	2.72	544	0.098	0.00006	0.5345 ± 1	0.065 ± 1	0.2227 ± 9	0.797 ± 15	58.8 ± 1.2	87	3,780 ± 55	4,153 ± 20	1,277 ± 25	4,339 ± 3
74-58	487	1,731	3.56	217	1.142	0.00071	0.5088 ± 2	0.012 ± 0	0.1419 ± 2	0.304 ± 6	21.4 ± 0.4	40	1,713 ± 29	3,155 ± 19	244 ± 6	4,267 ± 5
74-59	454	765	1.68	335	0.781	0.00049	0.5297 ± 1	0.048 ± 1	0.1637 ± 17	0.496 ± 10	36.2 ± 0.7	60	2,595 ± 41	3,672 ± 20	952 ± 21	4,326 ± 4
74-60	390	1,336	3.43	120	1.9	0.00119	0.5394 ± 3	0.010 ± 0	0.1641 ± 41	0.200 ± 4	14.9 ± 0.3	27	1,177 ± 21	2,808 ± 20	193 ± 6	4,353 ± 8
Session 2																
36-1	361	262	0.73	515	0.002	0.00001	0.5401 ± 8	0.232 ± 5	0.1743 ± 6	0.965 ± 18	71.9 ± 1.4	100	4,356 ± 60	4,355 ± 19	4,211 ± 74	4,355 ± 2
36-2	258	174	0.67	351	0.01	0.00006	0.5353 ± 10	0.226 ± 5	0.1641 ± 7	0.929 ± 18	68.6 ± 1.3	98	4,235 ± 59	4,307 ± 19	4,125 ± 75	4,341 ± 3
36-3	184	117	0.64	204	0	0	0.5143 ± 12	0.194 ± 4	0.1612 ± 10	0.768 ± 15	54.5 ± 1.1	86	3,674 ± 53	4,078 ± 20	3,588 ± 68	4,283 ± 4
36-4	250	188	0.75	333	0	0	0.5436 ± 10	0.218 ± 4	0.1831 ± 7	0.897 ± 17	67.2 ± 1.3	95	4,126 ± 58	4,287 ± 19	3,986 ± 72	4,364 ± 3
36-5	208	304	1.46	153	0	0	0.5163 ± 13	0.067 ± 1	0.1959 ± 10	0.498 ± 9	35.4 ± 0.7	61	2,604 ± 40	3,651 ± 19	1,306 ± 25	4,288 ± 4
36-6	346	940	2.71	182	0.001	0	0.5147 ± 11	0.026 ± 1	0.2005 ± 9	0.356 ± 7	25.2 ± 0.5	46	1,961 ± 32	3,317 ± 19	524 ± 10	4,284 ± 3
36-7	127	75	0.59	180	0	0	0.5587 ± 14	0.237 ± 5	0.1437 ± 9	0.968 ± 19	74.6 ± 1.5	99	4,364 ± 61	4,392 ± 20	4,291 ± 83	4,404 ± 4
36-8	329	339	1.03	337	0.013	0.00001	0.5272 ± 9	0.141 ± 3	0.2134 ± 8	0.683 ± 13	49.6 ± 0.9	78	3,355 ± 49	3,985 ± 19	2,672 ± 49	4,319 ± 3

* Common lead corrected using ^{204}Pb .

Data for zircon grain W74/2-36. f206% is $(\text{common } ^{206}\text{Pb}/\text{total } ^{206}\text{Pb}) \times 100$. Conc. (%) is percentage concordance defined as $[(^{206}\text{Pb}/^{238}\text{U})_{\text{age}}/(^{207}\text{Pb}/^{206}\text{Pb})_{\text{age}}] \times 100$. All errors are quoted at 1σ level. Errors given for Pb/Pb and Pb/Th ratios are based on counting statistics. Errors in Pb/U ratios also include an estimate of the Pb/U reproducibility error based on multiple analyses of the standard zircon C23 during the analytical sessions. During the first analytical session, all spots were given a unique number. In the second session, the grain was identified as W74-36 and each site was numbered consecutively, 36-1 to 36-8.

75–1,731 p.p.m., respectively (Table 1), reaching values considerably higher than those recorded from the ~4,276-Myr-old crystal⁴. There is no correlation between the $^{207}\text{Pb}/^{206}\text{Pb}$ age of the various analytical sites and the U, Th and Pb contents, or the Th/U ratio (Table 1), although the oldest site has the lowest values for all of these except Pb. These results suggest complex micrometre to sub-micrometre variations in composition, as indicated by several detailed studies^{18–20}, and also preclude any estimate of the composition of the host rock based on isotopic characteristics. The Th/U

ratio varies from 0.59 to 3.56, with those sites in the range 0.59 to 0.77 having $^{208}\text{Pb}/^{232}\text{Th}$ ages greater than 3,588 Myr and being the least discordant (Table 1). The large differences in Pb/U for similar $^{207}\text{Pb}/^{206}\text{Pb}$ ages (Table 1) indicate recent lead loss. This is especially evident within the combined light and dark rectangular area visible in the CL image, which may have formed part of the central region of the original zircon (Fig. 1b). The three spots from the second analytical run (36-3, 36-5 and 36-6), which partially overlap this zone, define a weighted $^{207}\text{Pb}/^{206}\text{Pb}$ age of $4,289 \pm 7$ Myr (2σ) that is younger than the $^{207}\text{Pb}/^{206}\text{Pb}$ age of all other sites (Table 1). A similar feature was noted in the core of a xenocrystic zircon older than 4.0 Gyr in granitic gneiss from Churla Well in the Narryer Terrane¹⁸ and was interpreted as reflecting an ancient episode of lead loss resulting from an event at high temperature. Unlike Churla Well, our data show no evidence of subsequent lead loss during Precambrian time, suggesting that the zircon was not affected by later igneous or metamorphic processes before its deposition in the conglomerate.

We interpret the concordant $^{207}\text{Pb}/^{206}\text{Pb}$ age of $4,404 \pm 8$ Myr (2σ) as recording the time of crystallization of zircon W74/2-36. Although this is represented by only one analysis, it is concordant (99%), is not affected by cracks, and there is no evidence from the analytical data that this site, rather than any other measured, may be anomalous. The crystal was subsequently affected by one or more episodes of ancient lead loss, but the timing of these is unresolved. The concordant and near-concordant ages of $4,364 \pm 6$ Myr,

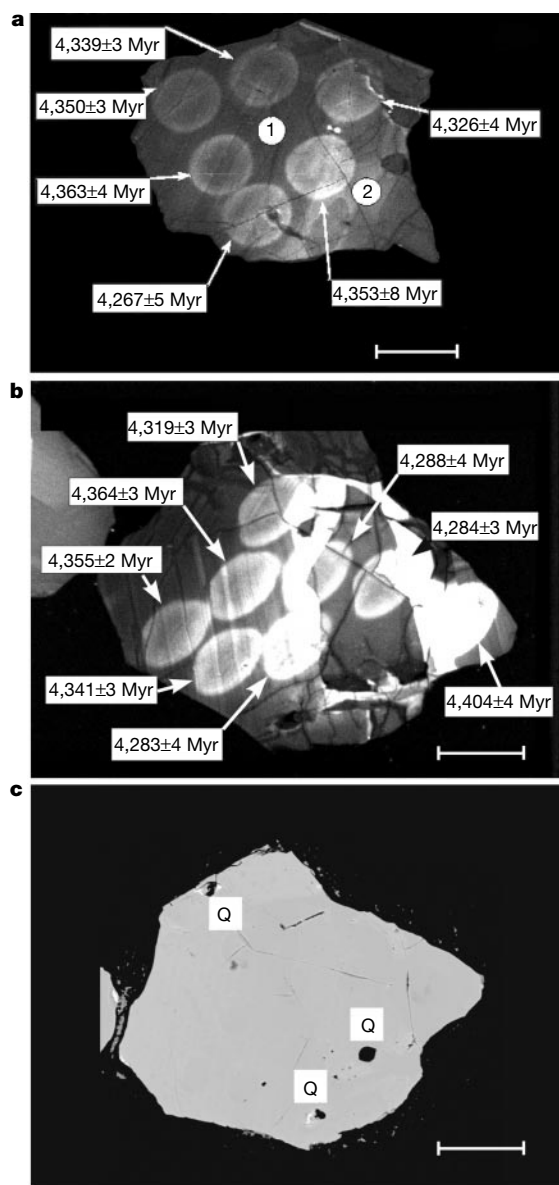


Figure 1 Cathodoluminescence and back-scattered electron images of zircon crystal W74/2-36. Scale bars are 50 μm . **a**, Image taken subsequent to first SHRIMP analysis. We note that lighter circular areas are the SHRIMP analytical sites and the values record the $^{207}\text{Pb}/^{206}\text{Pb}$ age (1σ) of each site. The two white areas represent the approximate location of the oxygen analytical spots, with $\delta^{18}\text{O}$ of 5‰ at point 1 and 7.4‰ at point 2. **b**, Cathodoluminescence image taken after the second SHRIMP analytical session showing the sites of analysis (as in **a**) and the $^{207}\text{Pb}/^{206}\text{Pb}$ ages (1σ) for each spot. We note the larger area showing light luminescence, the well defined cracks and the oscillatory zoning in both the light and dull portions of the crystal. **c**, Back-scattered electron image obtained after the second SHRIMP analytical session. We note the circular outline of the SHRIMP analytical sites and the lighter rectangular area, near the pointed termination, which corresponds to the dark area in the cathodoluminescence image (compare with **b**). The black areas (marked Q) are SiO_2 inclusions.

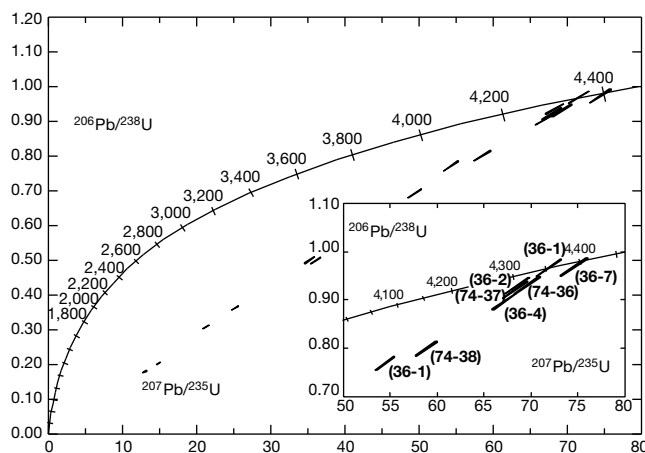


Figure 2 Combined concordia plot for grain W74/2-36, showing the U-Pb results obtained during the two analytical sessions. The inset shows the most concordant data points together with their analysis number (as in Table 1). Error boxes are shown at 1σ .

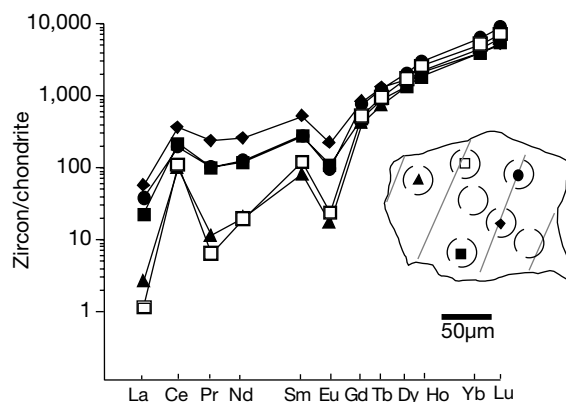


Figure 3 Rare earth element data for Jack Hills zircon W74-2/36 measured by ion microprobe. We note that spacing of elements along the x-axis is based on radii of 3^+ cations (in Å).

4,355 $\pm$ 4 Myr and 4,341 $\pm$ 6 Myr (2 σ) may represent actual geological events, possibly triggered by meteor bombardment¹⁸, or they may indicate slight changes in ²⁰⁷Pb/²⁰⁶Pb without generating much discordance², because of their position on the concordia curve (Fig. 2). The ²⁰⁷Pb/²⁰⁶Pb age of 4,289 $\pm$ 7 Myr (2 σ) for the three more discordant points from the light and dark area in the CL image (Fig. 1b) were most probably reset at this time by a high-temperature event, and subsequently underwent recent lead loss.

Most of the crystal exhibits weak banding in the CL images (Fig. 1a and b), indicating elemental variation marked by differences in luminescence which usually extend fully across the crystal, suggesting that the zircon was originally much larger. The development of banding parallel to the two sharp crystal boundaries in the top left-hand corner of the image and also within the bright area (Fig. 1b) suggests that it reflects oscillatory zoning of magmatic origin. The bright area almost entirely encloses the dark central region, which is devoid of any CL pattern.

Variation in the CL image is confirmed by the REE data (Fig. 3), which were obtained from the floors of five of the initial SHRIMP analytical pits. The analyses yielded prominent positive Ce anomalies, negative Eu anomalies, and heavy (H)REE-enrichment relative to light (L)REE contents; these are features typical of zircons analysed by ion microprobe¹¹ and from other Jack Hills detrital zircons¹³. The crystal is zoned with respect to LREE (La by about 10-fold), with greater LREE enrichment within and around the bright portion of the crystal. Calculated partition coefficients¹¹ yield magma compositions with marked LREE-enrichment for this part of the crystal (Lu/La_n (magma) is 238 to 616) and negative Eu anomalies, features indicative of evolved granitic melts. The crystal's growth in a silica-saturated environment is further supported by the identification of three areas, up to 20- μ m in diameter, with SiO₂ inclusions (Fig. 1c).

Four analyses of oxygen isotope ratio were made on the Edinburgh Cameca ion microprobe on two spots close to two of the initial SHRIMP sites. One spot was located near the 4,339 $\pm$ 3 Myr site in the low LREE region that is dull and banded in CL (Fig. 1a). This has a $\delta^{18}\text{O}$ of 5.0 $\pm$ 0.7‰ (1 σ), which is indistinguishable from the $\delta^{18}\text{O}$ values of mantle-derived zircon²¹ and zircon from juvenile, 3.0 to 2.7 Gyr tonalite-trondhjemite-granodiorite plutonic rocks of the Superior Province²². The second spot was located in the high LREE region of the grain, which is brighter in the CL image (Fig. 1a), close to the site yielding a ²⁰⁷Pb/²⁰⁶Pb age of 4,353 $\pm$ 8 Myr (2 σ). This has a $\delta^{18}\text{O}$ value of 7.4 $\pm$ 0.7‰ (1 σ), consistent with equilibrium with a magma of about 8.5 to 9.5‰. Magmatic oxygen isotope ratios in this range are indicative of involvement of supracrustal material which has undergone low-temperature interaction with a liquid hydrosphere; there is no known primitive or mantle reservoir of this composition.

REE and oxygen isotope compositions in grain W74/2-36 provide unique insights into the Earth's crust at 4.4 Gyr ago. They indicate formation in a granitic melt derived by partial melting of pre-existing crust. Calculated magma compositions show characteristics distinctive of evolved melts (such as LREE enrichment, prominent negative Eu anomalies, silica saturation), and are not consistent with the REEs of primitive, mafic rocks¹⁰. The variation from relatively high to low LREEs, coupled with a 2.4‰ decrease in the oxygen isotope ratio, is consistent with igneous zoning associated with melting and assimilation and subsequent equilibration with a mantle-derived melt. The initially high- $\delta^{18}\text{O}$ magma (~9‰) could result from melting of ocean crust that was hydrothermally altered at low temperatures²³, altered or evolved continental crust, or sediments. The high $\delta^{18}\text{O}$ value cannot result from closed-system fractional crystallization by a normal mantle melt. In the case of simple melting without assimilation, zircons in the melt would have a similar $\delta^{18}\text{O}$ value to zircons in its protolith. An alternative hypothesis, involving assimilation coupled with fractional crystallization (AFC) by a normal basalt of high $\delta^{18}\text{O}$ wallrock, is unlikely, as it would require either energetically unrealistic amounts of assimilation (>50%) or wallrocks with average $\delta^{18}\text{O}$ values above 12‰ which are uncommon even in the late Archaean²². The correlation of high $\delta^{18}\text{O}$ values and enriched LREEs is most consistent with the melting of continental crust or sediments, but whichever of these processes dominated, a large reservoir of liquid water is required on the surface of the Earth.

The zoning of zircon W74/2-36 provides further insights into magmatic processes at 4.4 Gyr ago. Such zones could form during AFC, but we suggest that the zircon first crystallized in a magma derived from the melting of existing continental crust at 4,404 $\pm$ 8 Myr ago, which subsequently formed a magmatic overgrowth when mixed with, or remelted by, a more primitive magma between 4,364 $\pm$ 6 Myr ago and 4,289 $\pm$ 7 Myr ago. This indicates that at 4.4 Gyr ago there were already intermediate to granitic, high- $\delta^{18}\text{O}$ continental rocks to contaminate the magma in which grain W74/2-36 grew.

The existence of liquid water at 4.4 Gyr ago could have important implications for the evolution of life. Microfossils as old as 3.5 Gyr are known²⁴. Metasediments and carbonaceous materials with biogenic carbon isotope ratios as low as $\delta^{13}\text{C} = -28\text{‰}$ are known at 3.8 Gyr ago²⁵. Zircon crystal W74/2-36 is over 500 Myr older than this organic matter and if liquid water was available to cause the evolved geochemistry that we have measured, then such water was also available for possible biological processes. High-energy asteroid bombardment before 3.9 Gyr ago is consistent with periodic formation and destruction of early oceans and the possibility that primitive life, if it evolved in the oceans, was globally extinguished more than once. □

Table 2 Rare earth element and oxygen isotope data

Trace element data (p.p.m.)												
Spot number	La	Ce	Pt	Nd	Sm	Eu	Gd	Tb	Dy	Ho	Yb	Lu
m2-13	5.5	134	9.4	55	41.4	6.3	100	34	338	106	666	135
m2-14	9.2	122	9.5	59	42.6	5.6	155	47	518	173	1065	225
m2-17	0.6	61	1.1	9	12.3	1.0	85	28	347	124	768	159
m2-30	13.6	226	21.8	119	78.4	12.7	170	49	424	119	672	130
m2-31	0.3	69	0.6	9	18.4	1.4	108	36	438	149	877	182
Oxygen isotope data												
Spot number	Analysis number		Measured $^{18}\text{O}/^{16}\text{O}$ ($\times 10^{-3}$)				$\delta^{18}\text{O}$ (VSMOW)					
m2-1	2(15)		1.9378				4.8					
m2-2d	2(16)		1.9388				5.3					
							Avg. 5.0					
m2-3	2(21)		1.9414				6.8					
m2-4d	2(22)		1.9436				8					
							Avg. 7.4					

Data were measured by ion microprobe for crystal W74-2/36.

Methods

Zircon data collection

Approximately 100 zircon grains were hand picked from a previously prepared +135- μm concentrate and mounted onto double-sided adhesive tape, along with pieces of the Curtin University Sri Lankan gem zircon standard (CZ3) with a conventionally-measured U-Pb age of 564 Myr (ref. 26). They were enclosed in epoxy resin disks, ground and polished so as to effectively cut all zircon grains in half, and then gold coated. Samples were imaged by cathodoluminescence (CL), resulting in a map that allowed us to identify grains, as well as providing information on internal structure that could be tested during analysis. During the first analytical session, a mass resolution of 5,400 was obtained and the error associated with the measurement of Pb/U isotopic ratios for the standard, at 1 standard deviation, was 1.96% for seven standards. After discovery of a grain with an age in excess of 4.3 Gyr, and following the collection of oxygen isotope and trace element data, the sample was reground, repolished and gold coated and, on the basis of a new CL image, the grain was reanalysed on SHRIMP II, with a total of eight new sites selected so that most were away from cracks in the crystal. During this session, the mass resolution was 4,885 and the error on the standard, at 1 standard deviation, was 1.86% for six standards. The relationship between measured Pb/U and UO/U ratios on SHRIMP follows a power-law equation with the exponent equal to two (ref. 27). The Pb/U ratios on the unknowns were normalized to those measured on the standard zircon (CZ3 – $^{206}\text{Pb}/^{238}\text{U} = 0.0914$). Both data sets were reduced following the methods of Nelson¹⁶, using the single-stage model Broken Hill common Pb correction for mass ^{204}Pb , since this is considered to be introduced chiefly through the gold coating¹⁶. The analytical spot size averaged 30 μm during each analytical run and each spot was rastered over 100 μm for five minutes before analysis to remove common Pb on the surface or contamination from the gold coating. All stated uncertainties and data listed in Table 1, and the error bars shown in Fig. 2, are at 1 σ ; ages discussed in the text are all 2 σ .

Oxygen isotope analysis

Four analyses were made by Cameca ims 4f ion microprobe on two spots on grain W74/2-36 using an energy offset of 350 eV. These were analysed for a total of 2×10^6 counts of ^{18}O for each analysis, yielding precision close to 0.7‰ (1 σ), based on gaussian counting statistics²⁸. Internal precision for each analysis was ± 0.3 and ± 0.6 ‰ (1 σ), comparing 80-cycle analysis halves on each full analysis, in agreement with theoretical counting statistics. The two point-analyses of the zircon crystal were interspersed with 11 analyses of Kim-5, a homogeneous zircon standard ($\delta^{18}\text{O} = 5.04 \pm 0.07$ ‰ VSMOW by laser fluorination) mounted in a separate standard block. Five standard zircons with 1.06 to 1.52 wt% HfO_2 were also analysed by ion microprobe and laser fluorination¹⁰, and no statistically significant dependence of instrumental mass fractionation on HfO_2 content was found²⁹ for the narrow range of Hf in these samples.

Rare earth element analysis

In situ determination of REEs was performed by Cameca ims 4f ion microprobe¹¹. Five ion microprobe REE analyses of the zircon crystal were made using a 14.5-keV primary beam of O^- defocused to an approximately 20 to 30 μm spot. Positive secondary ions were collected using an energy offset of 125 V. Analyses were standardized to the SRM-610 glass standard. Energy filtering and strategies to avoid and correct for isobaric interferences were used¹¹.

Received 1 August; accepted 24 November 2000.

- Ryder, G. Chronology of early bombardment in the inner solar system. *Geol. Soc. Am. Abstr. Progr.* 21, A299 (1992).
- Bowring, S. A. & Williams, I. S. Priscoan (4.00–4.03) orthogneisses from northwestern Canada. *Contrib. Mineral. Petrol.* **134**, 3–16 (1999).
- Froude, D. O. *et al.* Ion microprobe identification of 4,100–4,200 Myr-old terrestrial zircons. *Nature* **304**, 616–618 (1983).
- Compston, W. & Pidgeon, R. T. Jack Hills, evidence of more very old detrital zircons in Western Australia. *Nature* **321**, 766–769 (1986).
- Wilde, S. A. & Pidgeon, R. T. in *3rd International Archaeological Symposium (Perth), Excursion Guidebook* (eds Ho, S. E., Glover, J. E., Myers, J. S. & Muehlenbachs, K. R.) 82–95 (University of Western Australia Extension Publication, Vol. 21, Perth, 1990).
- Kober, B., Pidgeon, R. T. & Lippolt, H. J. Single-zircon dating by step-wise Pb-evaporation constrains the Archean history of detrital zircons from the Jack Hills, Western Australia. *Earth Planet. Sci. Lett.* **91**, 286–296 (1989).
- Amelin, Y. V. Geochronology of the Jack Hills detrital zircons by precise U-Pb isotope dilution analysis of crystal fragments. *Chem. Geol.* **146**, 25–38 (1998).
- Amelin, Y., Lee, D.-C., Halliday, A. N. & Pidgeon, R. T. Nature of the Earth's earliest crust from hafnium isotopes in single detrital zircons. *Nature* **399**, 252–255 (1999).
- Myers, J. S. in *Early Precambrian Processes* (eds Coward, M. P. & Ries, A. C.) 143–154 (Geological Society of London Special Publication No. 95, 1995).
- Peck, W. H., Valley, J. W., Wilde, S. A. & Graham, C. M. Oxygen isotope ratios and rare earth elements in 3.3 to >4.0 Ga zircons: ion microprobe evidence for Early Archean high $\delta^{18}\text{O}$ continental crust. *Geochim. Cosmochim. Acta* (submitted).
- Hinton, R. W. & Upton, B. G. J. The chemistry of zircon; variations within and between large crystals from syenite and alkali basalt xenoliths. *Geochim. Cosmochim. Acta* **55**, 3287–3302 (1991).
- Valley, J. W., Chiarenzelli, J. R. & McLelland, J. M. Oxygen isotope geochemistry of zircon. *Earth Planet. Sci. Lett.* **126**, 187–206 (1994).
- Maas, R., Kinny, P. D., Williams, I. S., Froude, D. O. & Compston, W. The earth's oldest known crust: a geochronological and geochemical study of 3900–4200 Ma old detrital zircons from Mt. Narryer and Jack Hills, Western Australia. *Geochim. Cosmochim. Acta* **56**, 1281–1300 (1992).
- Compston, W., Williams, I. S., Kirschvink, J. L., Zhang, Z. & Ma, G. Zircon U-Pb ages for the Early

- Cambrian time-scale. *J. Geol. Soc. Lond.* **149**, 171–184 (1992).
- Williams, I. S. in *Applications of Microanalytical Techniques to Understanding Mineralizing Processes* (eds McKibben, M. A., Shanks III, W. C. & Ridley, W. I.) 1–95 (Reviews in Economic Geology, Vol. 7, Society of Economic Geologists, Littleton, Colorado, 1998).
- Nelson, D. R. Compilation of SHRIMP U-Pb geochronology data, 1996. *Geol. Surv. Western Australia Rec.* **1997/2**, 1–11 (1997).
- Taylor, S. R. *Solar System Evolution: A New Perspective* 289 (Cambridge Univ. Press, Cambridge, 1992).
- Nelson, D. R., Robinson, B. W. & Myers, J. S. Complex geological histories extending from >4.0 Ga deciphered from xenocryst zircon microstructures. *Earth Planet. Sci. Lett.* **181**, 89–102 (2000).
- Williams, I. S., Compston, W., Black, L. P., Ireland, T. R. & Foster, J. J. Unsupported radiogenic Pb in zircon: a cause of anomalously high Pb-Pb, U-Pb and Th-Pb ages. *Contrib. Mineral. Petrol.* **88**, 322–327 (1984).
- Mattinson, J. M. A study of complex discordance in zircons using step-wise dissolution techniques. *Contrib. Mineral. Petrol.* **16**, 117–129 (1994).
- Valley, J. W., Kinny, P. D., Schulze, D. J. & Spicuzza, M. J. Zircon megacrysts from kimberlite: oxygen isotope variability among mantle melts. *Contrib. Mineral. Petrol.* **133**, 1–11 (1998).
- King, E. M., Valley, J. W., Davis, D. W. & Edwards, G. R. Oxygen isotope ratios of Archean plutonic zircons from granite-greenstone belts of the Superior province: indicator of magmatic source. *Precamb. Res.* **92**, 365–387 (1998).
- Muehlenbachs, K. in *Stable Isotopes* (eds Valley, J. W. *et al.*) *MSA Rev. Min.* **16**, 425–444 (1986).
- Schopf, J. W. Microfossils in the early Archean Apex Chert: New evidence for the antiquity of life. *Science* **260**, 640–646 (1993).
- Hayes, J. M., Kaplan, I. R. & Wedeking, K. W. in *Earth's Earliest Biosphere; its Origin and evolution* 93–134 (Princeton Univ. Press, Princeton, NJ, 1983).
- Pidgeon, R. T. *et al.* in *Eighth Int. Conf. Geochron., Cosmochron. Isotope Geol.* (eds Lanphere, M. A., Dalrymple, G. B. & Turrin, B. D.) 251 (US Geological Survey Circular 1107, Denver, Colorado, 1994).
- Claué-Long, J. C., Compston, W., Roberts, J. & Fanning, C. M. in *Geochronology, Time Scales and Global Stratigraphic Correlation* (eds Berggren, W. A., Kent, D. V., Aubry, M.-P. & Hardenbol, J.) 3–21 (Soc. of Sedimentary Geology, SEPM Sp. Publ. 4, 1995).
- Valley, J. W., Graham, C. M., Harte, B., Eiler, J. M. & Kinny, P. D. in *Applications of Microanalysis to Understanding Mineralizing Processes* (eds McKibben, M. A., Shanks III, W. C. & Ridley, W. I.) 73–98 (Reviews in Economic Geology, Vol. 7, Society of Economic Geologists, Littleton, Colorado, 1998).
- Eiler, J. M., Graham, C. M. & Valley, J. W. SIMS analysis of oxygen isotopes: matrix effects in complex minerals and glasses. *Chem. Geol.* **138**, 221–244 (1997).

Acknowledgements

We thank A. Nemchin for assistance with the cathodoluminescence imaging, J. Craven for assistance in stable isotope analysis by ion microprobe and J. Fournelle for assistance with electron microprobe analysis. Initial fieldwork was supported by the Australian Research Council and analytical work by NERC, NSF and the US Department of Energy. D. Nelson and K. McNamara kindly commented on the manuscript.

Correspondence and requests for materials should be addressed to S.A.W. (e-mail: wildes@lithos.curtin.edu.au).

Oxygen-isotope evidence from ancient zircons for liquid water at the Earth's surface 4,300 Myr ago

Stephen J. Mojzsis^{*†}, T. Mark Harrison^{*} & Robert T. Pidgeon[‡]

^{*} Department of Earth and Space Sciences and IGPP Center for Astrobiology, University of California Los Angeles, Los Angeles, California 90095-1567, USA
[‡] Department of Applied Geology, Curtin University of Technology, Perth, WA 6001, Australia

Granitoid gneisses and supracrustal rocks that are 3,800–4,000 Myr old are the oldest recognized exposures of continental crust¹. To obtain insight into conditions at the Earth's surface more than 4 Gyr ago requires the analysis of yet older rocks or their mineral remnants. Such an opportunity is presented by detrital zircons more than 4 Gyr old found within 3-Gyr-old quartzitic rocks in the Murchison District of Western Australia^{2,3}. Here we report *in situ* U–Pb and oxygen isotope results for such zircons that place constraints on the age and composition of their

[†] Present address: Department of Geological Sciences, University of Colorado, Boulder, Colorado 80309-0399, USA.

sources and may therefore provide information about the nature of the Earth's early surface. We find that 3,910–4,280 Myr old zircons have oxygen isotope ($\delta^{18}\text{O}$) values ranging from $5.4 \pm 0.6\text{‰}$ to $15.0 \pm 0.4\text{‰}$. On the basis of these results, we postulate that the $\sim 4,300$ -Myr-old zircons formed from magmas containing a significant component of re-worked continental crust that formed in the presence of water near the Earth's surface. These data are therefore consistent with the presence of a hydrosphere interacting with the crust by 4,300 Myr ago.

The Narryer Gneiss Complex, Western Australia, contains 3.73–3.30-Gyr-old granitic to tonalitic gneisses and about 3.0-Gyr-old metasedimentary rocks in the Mt Narryer and Jack Hills regions (Fig. 1). Previously, detrital zircons ranging in age from 4.27 to 3.05 Gyr (refs 2–5) have been described from these two regions. The source rocks of these zircons have not been identified and may not have been preserved.

Because zircon is a common accessory phase in granitoids and their volcanic equivalents, these rocks are generally considered to be the dominant source of detrital zircons. But zircon can form in other rock types (albeit of low abundance in most cases) such as syenites, carbonatites, and mafic rocks⁶, and can also grow in hydrothermal veins and during metamorphism. Residual plagiogranites from Iceland-type melts more than 4 Gyr old within a predominantly mafic crust have been suggested as the source of the oldest detrital zircons from the Narryer Gneiss Complex⁷, but this view has been challenged on trace-element and mineralogical grounds⁸.

Zircon has been extensively studied because it provides reliable U–Pb crystallization ages and can survive both high-grade metamorphism and sedimentary transport processes. Measurements of oxygen isotopes in rocks and minerals help us to understand the magmatic, fluid and thermal history of the crust. Exchange rates for oxygen are very slow⁹ in zircon, which may permit preservation of the protolith oxygen isotope composition through high-grade metamorphism¹⁰. Oxygen isotopes are used to discriminate among possible granitoid sources (mantle, metasedimentary, hybrid); thus, measurements of ancient zircons from the Narryer Gneiss Complex could allow us to directly explore crust–hydrosphere interactions before 4 Gyr ago.

The zircon-bearing rocks in the Jack Hills form part of a thick (>2 km) series of fan delta sequences deposited in a fault-bounded cratonic margin¹¹ that were subsequently metamorphosed to upper greenschist facies^{8,13}. At least three distinct age groupings of ancient zircons have been described^{3,8,11,13} (~ 4.3 Gyr, 4.15 Gyr and 3.9 Gyr) that could represent discrete source terranes, although no rocks

older than ~ 3.75 Gyr have been found in the region^{4,5,12}. We collected quartz-pebble conglomerates rich in heavy mineral indicators (such as Cr-spinel, fuchsite) from a locality in the Erawondoo region of the Jack Hills (Fig. 1) that was previously sampled for investigations of the trace element composition and inclusion mineralogy of detrital zircons more than 4 Gyr old^{2,3,8,13}. In order to minimize the loss of zircon during sample preparation, we crushed, powdered and sieved 15 kg of sample JH992 and processed all fractions through heavy liquids without pre-washing. This yielded about 80 zircons kg^{-1} in the size range 50–700 μm , most with low aspect ratios. For each sample mount, we placed 100 zircon grains on adhesive tape together with standard zircon AS3¹⁴ and filled an enclosing one-inch diameter mould with epoxy. After curing, the mount was cleaned and polished following our usual procedures¹⁵ and the internal features of the zircons were imaged using optical and back-scattered electron microscopy.

Because the very oldest grains are less than 10% of the total zircon population, we developed a new multi-collector ion microprobe technique to rapidly identify the oldest grains by determining $^{207}\text{Pb}/^{206}\text{Pb}$ ages by simultaneous measurement of ^{207}Pb and ^{206}Pb on adjacent electron multipliers. Using a primary O^+ beam of 3 nA focused to an approximately 20 μm diameter spot together with O_2 flooding¹⁶, this new approach yielded precise (typically ± 10 Myr) $^{207}\text{Pb}/^{206}\text{Pb}$ ages in 3.5 minutes, permitting us quickly to survey 200 zircons. The oldest candidate grains were then targeted for conventional U–Pb ion microprobe measurements. Our search revealed 17 zircons more than 3.9 Gyr old. Grains JH992_95 and JH992_48 (Table 1) showed very ancient and essentially concordant ages of $4,279 \pm 5$ Myr (2σ) and $4,280 \pm 5$ Myr (2σ), respectively, suggesting that $\sim 1\%$ of the zircons in our sample are about 4.3 Gyr old.

We used a multi-collector method to obtain precise oxygen isotope compositions of microscale domains within the zircons. This new method provides the spatial resolution to determine isotopic heterogeneity at the sub-grain scale and avoid metamict zones or later metamorphic overgrowths that might complicate interpretation of the isotopic data. After the U–Pb measurements, the zircon mounts were re-polished, cleaned, and gold-coated and then analysed for oxygen isotopes using a 6-nA Cs^+ primary beam focused to a 25 μm diameter spot. In most cases, the oxygen isotope measurement was made directly adjacent to the spot on which a U–Pb age was determined. A mass resolving power of $\sim 2,000$ was sufficient to separate $^{18}\text{O}^-$ and $^{16}\text{O}^-$ from molecular interferences. Each beam was focused into a Faraday cup with a typical count rate

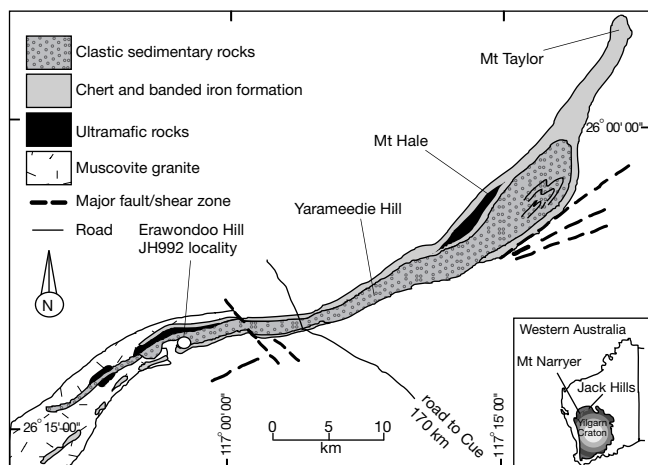


Figure 1 Schematic geological map of the Erawondoo region, Western Australia, showing the location of quartz-pebble conglomerate sample JH992 containing detrital zircons more than 4 Gyr old. The Jack Hills and the Mt Narryer quartzitic units are part of an arcuate metasedimentary belt near the northwestern margin of the Yilgarn Craton (see inset).

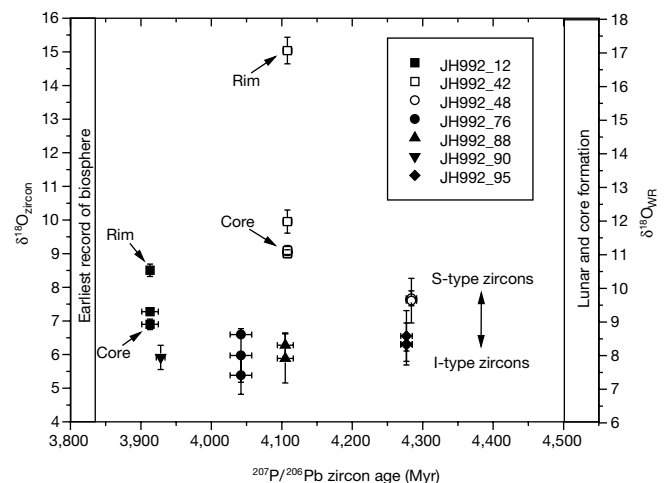


Figure 2 Ion microprobe $\delta^{18}\text{O}$ data for individual zircon spot analyses versus $^{207}\text{Pb}/^{206}\text{Pb}$ zircon age (Supplementary information available at <http://www.nature.com>). The right vertical axis shows the estimated $\delta^{18}\text{O}$ data for the whole rock^{10,25} ($\delta^{18}\text{O}_{\text{WR}}$) from which the zircon crystallized. High $\delta^{18}\text{O}_{\text{WR}}$ values are consistent with the incorporation of recycled crustal material that had interacted with low-temperature water in the magmatic source of the zircons.

of 2 GHz for $^{16}\text{O}^-$ and 4 MHz for $^{18}\text{O}^-$; the total integration time for each analysis was less than 6 min. Backgrounds on the Faraday cup system were monitored periodically and the uncertainty in this correction was always less than 0.1‰. Instrumental mass fractionation was corrected by reference to standard zircons KIM5 ($5.04 \pm 0.05\text{‰}$)¹⁷ and 91500 ($9.8 \pm 0.2\text{‰}$; S. Claesson, personal communication), which were interposed between analyses of unknowns. Daily reproducibility determined from more than 15 spots was typically better than $\pm 0.2\text{‰}$ for both standards. Uncertainties in oxygen isotope data are reported at the 2σ level. Most zircons were examined for intracrystal homogeneity by analysing multiple spots. Two zircons (JH992_12, and JH992_42) were found to have core to rim variability over about 50 μm that correlates with observed overgrowths (Table 1), emphasizing our need for the highest possible spatial resolution. The values of $\delta^{18}\text{O}$ for 3.90–4.28-Gyr-old zircon cores ranged from $5.4 \pm 0.6\text{‰}$ (JH992_76) to $9.0 \pm 0.2\text{‰}$ (JH992_42) for seven zircons from sample JH992. Figure 2 shows the data for these zircons plotted as $\delta^{18}\text{O}_{\text{zircon}}$ versus $^{207}\text{Pb}/^{206}\text{Pb}$ age.

The $\delta^{18}\text{O}_{\text{SMOW}}$ of the mantle, estimated to be 5.5‰ ¹⁸, is unlikely to have changed over time as contemporary mid-ocean ridge and ocean island basalts have identical $\delta^{18}\text{O}$ values of $5.7 \pm 0.2\text{‰}$ (refs 19, 20). Crustal contamination can strongly affect oxygen isotopes in igneous rocks. Phanerozoic granitoids derived largely from orthogneiss protoliths (I-types) tend to have $\delta^{18}\text{O}$ values below 9‰, whereas those derived by melting of clay-rich sedimentary rocks (S-types) have higher $\delta^{18}\text{O}$ values²¹. Granitoids with $\delta^{18}\text{O}$ values significantly less than 6‰ probably reflect interaction with meteoric water²². In general, S-type granitoids form by partial melting of metasediments enriched in ^{18}O , as opposed to I-type granitoids, which form by melting of igneous rocks derived from arc processes²³. As the average $\delta^{18}\text{O}$ values of Archean sedimentary rocks varies from 9 to 12‰ (ref. 24), only small inputs of this component to a primitive source magma will measurably raise its $\delta^{18}\text{O}$.

The $\delta^{18}\text{O}$ values of the zircons we examined range from 5.4‰ to 15.0‰ (Table 1; Fig. 2). In the two most discordant grains (JH992_12, JH992_42), we identified core–rim relationships and analysed both. In these cases, rim compositions were systematically higher in $\delta^{18}\text{O}$ than cores. In the five other zircons investigated, multiple analyses undertaken on optically continuous grains range in $\delta^{18}\text{O}$ from 5.4 to 7.7‰. The oxygen isotope fractionation between zircon and a granitoid host rock is approximately -2‰ (refs 10, 25), permitting us to estimate the $\delta^{18}\text{O}$ value of the melt from which the zircon crystallized. Source $\delta^{18}\text{O}$ values calculated in this fashion for all seven zircon cores analyzed range from about 7 to

11‰. These results indicate the presence in these zircons of recycled crustal material that had interacted with liquid water under surface, or near-surface, conditions. This conclusion is consistent with the results of a Hf isotope study of detrital zircons from the Narryer Gneiss Complex²⁶ that shows that many Hadean zircons formed by re-melting of significantly older crust. If the rims of grains JH992_12 and JH992_42 also reflect crystallization from a melt, then a source $\delta^{18}\text{O}$ value as high as 17‰ is possible.

We note that mineral assemblages (including muscovite and monazite) suggestive of derivation from a peraluminous melt were found as inclusions in 4.2-Gyr-old zircons from the Narryer Gneiss Complex⁸. Peraluminous granitoids form mainly from melting of a graywacke protolith²⁷; therefore, this observation is also consistent with the presence of a hydrosphere on Earth before 4.2 Gyr ago.

The crystallization of Hadean zircons containing heavy oxygen isotopic compositions suggests the existence of a hydrosphere at or near the Earth's surface within about 200 Myr of terrestrial core formation and the origin of the Moon at 4.50 Gyr (ref. 28). Liquid water, a source of energy, and organic raw materials are assumed to be necessary for the origin and propagation of life. Our results thus raise the possibility that a biosphere could have arisen on Earth at least 400 Myr earlier than is now thought²⁹. □

Received 26 September; accepted 1 December 2000.

- Goodwin, A. M. *Principles of Precambrian Geology* (Academic, New York, 1996).
- Froude, D. O. *et al.* Ion microprobe identification of 4100–4200 Myr old terrestrial zircons. *Nature* **304**, 616–618 (1983).
- Compston, W. & Pidgeon, R. T. Jack Hills, evidence of more very old detrital zircons in Western Australia. *Nature* **321**, 766–769 (1986).
- Nutman, A. P., Kinny, P. D., Compston, W. & Williams, I. S. SHRIMP U–Pb zircon geochronology of the Narryer Gneiss Complex, Western Australia. *Precambrian Res.* **52**, 275–300 (1991).
- Nelson, D. R., Robinson, B. W. & Myers, J. S. Complex geological histories extending for ≥ 4.0 Ga deciphered from xenocryst zircon microstructures. *Earth Planet. Sci. Lett.* **181**, 89–102 (2000).
- DeLong, S. E. & Chatelain, C. Trace-element constraints on accessory-phase saturation in evolved MORB magma. *Earth Planet. Sci. Lett.* **101**, 206–215 (1990).
- Galer, S. J. G. & Goldstein, S. L. Early mantle differentiation and its thermal consequences. *Geochim. Cosmochim. Acta* **55**, 227–239 (1991).
- Maas, R., Kinny, P. D., Williams, I. S., Froude, D. O. & Compston, W. The Earth's earliest known crust: A geochronological and geochemical study of 3900–4200 Ma old zircons from Mt. Narryer and Jack Hills, Western Australia. *Geochim. Cosmochim. Acta* **56**, 1281–1300 (1992).
- Watson, E. B. & Cherniak, D. J. Oxygen diffusion in zircon. *Earth Planet. Sci. Lett.* **148**, 527–544 (1997).
- Valley, J. W., Chiarenzelli, J. R. & McLelland, J. M. Oxygen isotope geochemistry of zircon. *Earth Planet. Sci. Lett.* **126**, 187–206 (1994).
- Myers, J. S. & Williams, I. R. Early Precambrian crustal evolution at Mt. Narryer, Western Australia. *Precambrian Res.* **27**, 153–163 (1985).
- Myers, J. S. Early Archean Narryer Gneiss Complex, Yilgarn Craton, Western Australia. *Precambrian Res.* **38**, 297–307 (1988).
- Maas, R. & McCulloch, M. T. The provenance of Archean clastic metasediments in the Narryer Gneiss Complex, Western Australia: Trace element geochemistry, Nd isotopes, and U–Pb ages for

Table 1 U–Pb geochronology and oxygen isotope compositions

Grain spot U–Pb	$\%^{206}\text{Pb}^*$ (%)	$^{206}\text{Pb}/^{238}\text{Pb}$ age (Myr)	$^{207}\text{Pb}/^{235}\text{U}$ age (Myr)	$^{207}\text{Pb}/^{206}\text{Pb}$ age (Myr)	Th/U	$\delta^{18}\text{O}_{\text{zircon}}$ (‰)
12_1	98.9	780 ± 97	2126 ± 118	3913 ± 12	2.6	7.3 ± 0.1 6.9 ± 0.2 $8.5 \pm 0.2^\dagger$
42_1	99.0	657 ± 97	2081 ± 137	4108 ± 6	2.4	$10.0 \pm 0.3^\dagger$ 9.0 ± 0.1 9.0 ± 0.2 $15.0 \pm 0.4^\dagger$
48_1	99.9	3582 ± 209	4045 ± 76	4284 ± 7	0.92	7.7 ± 0.2
48_2	99.9	3970 ± 101	4175 ± 33	4276 ± 6	0.88	7.6 ± 0.7
48_3	99.9	4147 ± 167	4238 ± 65	4282 ± 6	0.87	
76_1	99.7	3567 ± 288	3877 ± 106	4042 ± 15	0.30	6.6 ± 0.1 6.0 ± 0.8 5.4 ± 0.6
88_1	99.9	3928 ± 237	4046 ± 78	4105 ± 11	0.52	5.9 ± 0.7 6.3 ± 0.4
90_1	98.6	1345 ± 132	2676 ± 102	3928 ± 6	3.1	5.9 ± 0.4
95_1	99.9	4284 ± 143	4283 ± 46	4282 ± 7	0.31	6.3 ± 0.2
95_2	99.9	3730 ± 236	4093 ± 81	4277 ± 8	0.44	6.6 ± 0.8 6.3 ± 0.6

Data is from detrital zircons that are more than 3.9 Gyr old from sample JH992, Jack Hills, Western Australia.

Uncertainties in ages are reported at 1σ . Oxygen isotope values are expressed in per mil deviations from standard mean ocean water (SMOW); uncertainties in $\delta^{18}\text{O}$ are reported at 2σ .

* Radiogenic lead.

$^\dagger \delta^{18}\text{O}_{\text{zircon}}$ data measured at the rim of the zircon; all others measured at the core.

- detrital zircons. *Geochim. Cosmochim. Acta* **55**, 1915–1932 (1991).
14. Paces, J. B. & Miller, J. D. Precise U–Pb ages of Duluth Complex and related mafic intrusions, northeastern Minnesota—geochronological insights to physical, petrogenetic, paleomagnetic and tectonomagmatic processes associated with the 1.1 Ga midcontinent rift system. *J. Geophys. Res.* **98**, 13997–14013 (1993).
 15. Quiddeleur, X. *et al.* Thermal evolution and slip history of the Renbu Zedong Thrust, southeastern Tibet. *J. Geophys. Res.* **102**, 2659–2679 (1997).
 16. Schuhmacher, M., de Chambost, E., McKeegan, K. D., Harrison, T. M. & Migeon, H. in *Secondary Ion Mass Spectrometry SIMS IX* (eds Benninghoven, A. *et al.*) 919–922 (1994).
 17. Valley, J. W., Kinny, P. D., Schulze, D. J. & Spicuzza, M. J. Zircon megacrysts from kimberlite: oxygen isotope variability among mantle melts. *Contrib. Mineral. Petrol.* **133**, 1–11 (1998).
 18. Matthey, D., Lowry, D. & McPherson, C. Oxygen isotope composition of mantle peridotite. *Earth Planet. Sci. Lett.* **128**, 231–241 (1994).
 19. Harmon, R. S. & Hoefs, J. Oxygen isotope heterogeneity of the mantle deduced from global ^{18}O systematics of basalts from different geotectonic settings. *Contrib. Mineral. Petrol.* **120**, 95–114 (1995).
 20. Eiler, J. M. *et al.* Oxygen isotope variations in ocean island basalt phenocrysts. *Geochim. Cosmochim. Acta* **61**, 2281–2293 (1997).
 21. O’Neil, J. R. & Chappell, B. W. Oxygen and hydrogen isotope variations in the Berridale Batholith, Southeastern Australia. *J. Geol. Soc. Lond.* **133**, 559–571 (1977).
 22. Taylor, H. P. & Sheppard, S. M. F. in *Stable Isotopes in High Temperature Processes* (eds Valley, J. W. *et al.*) *Rev. Mineral.* **16**, 227–271 (Mineralogical Society of America, Washington DC, 1986).
 23. Chivas, A. R., Andrew, A. S., Sinha, A. K. & O’Neill, J. R. Geochemistry of Pliocene–Pleistocene oceanic arc plutonic complex, Guadalcanal. *Nature* **300**, 139–143 (1982).
 24. Longstaffe, F. J. & Schwarcz, H. P. $^{18}\text{O}/^{16}\text{O}$ of Archean clastic metasedimentary rocks: A petrogenetic indicator for Archean gneisses? *Geochim. Cosmochim. Acta* **41**, 1303–1312 (1977).
 25. Zheng, Y.-F. Calculation of oxygen isotope fractionation in anhydrous silicate minerals. *Geochim. Cosmochim. Acta* **57**, 1079–1091 (1993).
 26. Amelin, Y., Lee, D. C., Halliday, A. N. & Pidgeon, R. T. Nature of the Earth’s earliest crust from hafnium isotopes in single detrital zircons. *Nature* **399**, 252–255 (1999).
 27. White, A. J. R. & Chappell, B. W. Ultrametamorphism and granitoid genesis. *Tectonophysics* **43**, 7–22 (1977).
 28. Halliday, A. Terrestrial accretion rates and the origin of the Moon. *Earth Planet. Sci. Lett.* **176**, 17–30 (2000).
 29. Mojzsis, S. J. *et al.* Evidence for life on Earth before 3,800 million years ago. *Nature* **384**, 55–59 (1996).

Supplementary information is available on Nature’s World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

This work was supported by grants from NASA and NSF. Technical assistance from C. D. Coath is gratefully appreciated. We thank S. Claesson and J. Valley for providing the zircon oxygen standards. We also thank A. Halliday and C. Miller for comments on the manuscript. We are grateful to the McTaggart family of Mt Narryer station and the Broad family of Milly Milly station, Western Australia, for their hospitality in the field. We acknowledge support from the Instrumentation and Facilities Program of the National Science Foundation.

Correspondence and requests for materials should be addressed to S.J.M. (e-mail: Steve.Mojzsis@lasp.colorado.edu).

Fossil that fills a critical gap in avian evolution

Mark A. Norell* & Julia A. Clarke†

* Division of Paleontology, American Museum of Natural History, 79th Street at Central Park West, New York, New York 10024-5192, USA

† Department of Geology and Geophysics, Yale University, PO Box 208109, New Haven, Connecticut 06520-8109, USA

Despite the discoveries of well-preserved Mesozoic birds^{1–5}, a key part of avian evolution, close to the radiation of all living birds (Aves), remains poorly represented⁶. Here we report on a new taxon from the Late Cretaceous locality of Ukhaa Tolgod, Mongolia⁷, that offers insight into this critically unsampled period. *Apsaravis* and the controversial alvarezsaurids⁸ are the only avialan⁹ taxa known from the continental deposits at Ukhaa Tolgod, which have produced hundreds of fossil mammals, lizards and other small dinosaurs⁷. The new taxon, *Apsaravis ukhaana*, is the best-preserved specimen of a Mesozoic ornithurine bird discovered in over a century. It provides data important for assessing morphological evolution across Avialae, with implications for,

first, the monophyly of Enantiornithes and Sauriurae; second, the proposition that the Mesozoic sister taxa of extant birds, as part of an ‘ecological bottleneck’, inhabited exclusively near-shore and marine environments²; and third, the evolution of flight after its origin.

Systematic paleontology

Theropoda Marsh 1881

Avialae Gauthier 1986

Ornithurae Chiappe 1991

Apsaravis ukhaana new taxon

Type specimen. IGM (Institute of Geology, Mongolia) 100/1017 (Figs 1–3).

Etymology. ‘*Apsara*’ (sanskrit), winged consorts prominent in Buddhist and Hindu art, plus ‘avis’ from the Greek *aves* for bird. ‘Ukhaana’ refers to the type locality Ukhaa Tolgod.

Type locality. Camels Humps amphitheater, Ukhaa Tolgod, Mongolia, ?Campanian⁷.

Diagnosis. *Apsaravis* is differentiated from other avialans by the derived presence of a strong tubercle on the proximal posterior surface of the humerus directly distal to the humeral head (Fig. 1), a hypertrophied trochanteric crest on the femur and well-projected wings of the sulcus cartilaginis tibialis on the posterodistal tibiotarsus (Figs 1 and 3). Nine additional characters optimize as local apomorphies in the phylogenetic analysis (Fig. 4).

Description. The holotype specimen is three-dimensionally preserved and partially articulated (Figs 1 and 2). The fragmentary skull has a ring of scleral ossicles in the left orbit. A fragment (the left quadrate?), lies against the cranium close to the first of 12 preserved heterocoelic cervical vertebrae (Fig. 2). Part of the right jaw is a crushed arch of bone near the first phalanx of right manual digit II (Fig. 1). Another fragment includes the tips of toothless dentaries joined in an ossified symphysis (Fig. 1).

Seven opisthocoelic dorsal vertebrae and ten ankylosed sacral vertebrae are visible ventrally. Five free caudal vertebrae and a pygostyle, broken at its base and distal tip, complete the series. The pectoral and pelvic arches are nearly complete. The V-shaped ulnae and radia are preserved and metacarpals I–III are fused (at least proximally) into a carpometacarpus (Figs 1 and 2). The proximal tarsals and tibia are coossified (Figs 1 and 3). The right tibiotarsus is in articulation with the tarsometatarsus (Fig. 3). In both tarsometatarsi, the distal tarsals and metatarsals II–IV are co-ossified completely and enclose the distal vascular foramen. Metatarsal V is not present, and pedal digit I was not present or is not preserved. Distal to the area of the hypotarsus, an ossified tendon lies against metatarsal III. Semi-articulated pedal phalanges (Figs 1 and 3) decrease in length distally.

The phylogenetic position of *Apsaravis* close to crown-clade Aves (Fig. 4) is strongly supported. *Apsaravis* uniquely and unambiguously shares with Aves, and the only other well-preserved Mesozoic ornithurines¹ *Ichthyornis* and *Hesperornithiformes*¹⁰ (for example, *Hesperornis* and *Baptornis*), at least ten ankylosed sacral vertebrae; pubis and ischium sub-parallel and closely appressed; pubis medio-laterally compressed; a patellar groove on the distal femur; lateral condyle of the tibiotarsus equal to, or surpassing, the width of the medial; and metatarsal III pinched proximally between II and IV.

Other derived forelimb characters present in *Apsaravis*, *Ichthyornis* and Aves are indeterminable in the highly apomorphic, diving *Hesperornithiformes*. These include a globose humeral head; an extensor process on the proximal surface of metacarpal I; phalanx 1 of manual digit II posteriorly compressed; and a flat, and posteriorly bowed, metacarpal III less than one-half the diameter of metacarpal II. *Apsaravis* shares with *Ichthyornis* and Aves the derived presence of an anteriorly developed midline ridge associated with an anterior sternal keel, which is absent in flightless *Hesperornithiformes*, missing data in *Patagopteryx deferrariisi*¹¹ and unknown in more basal avialans where a keel, if present, is posteriorly developed^{5,12}. *Apsaravis*, *Ichthyornis* and Aves also share the derived

presence of an 'obturator flange' on the proximal ischium and a mediolaterally orientated postacetabular ilium (Fig. 1); both characters are not present in Hesperornithiformes or more basal taxa.

The *Apsaravis* hypotarsus lacks distinct ridges or grooves as in Hesperornithiformes and *Patagopteryx*¹¹. *Ichthyornis* shares with Aves, exclusive of *Apsaravis* and more basal avialans, the presence of at least one hypotarsal ridge, a procoracoid process and a medially curved acrocoracoid process. *Apsaravis* and *Ichthyornis* lack synapomorphies of Aves: the deltopectoral crest of the humerus projects primitively dorsally not anteriorly; the humeral pneumatic foramen does not contain a pneumatic foramen; there is one proximal vascular foramen in the proximal tarsometatarsus as opposed to two; and an ossified supratendinal bridge on the tibiotarsus is absent.

Several characters shared with *Patagopteryx* and Ornithurae place

Apsaravis closer to Aves than Enantiornithes. These taxa share the derived presence of completely heterocoelic cervical vertebrae (absent in *Ichthyornis*); a curved scapular shaft; loss of the cupped muscle fossa on the ilium; distally non-contacting pubes; metatarsals and distal tarsals fused to form a tarsometatarsus with the distal vascular foramen enclosed; a discrete area on the posteroproximal metatarsals for an incipient hypotarsus; and a proximal vascular foramen between metatarsals III and IV (absent in Hesperornithiformes). Twenty-seven unambiguous synapomorphies place *Apsaravis* as a derived ornithurine (Fig. 4).

Apsaravis displays several previously reported enantiornithine synapomorphies^{11–16} that our analysis shows to either be plesiomorphic or no longer optimize as unique to Enantiornithes because they also occur in *Apsaravis* and/or other avialan taxa. For example,

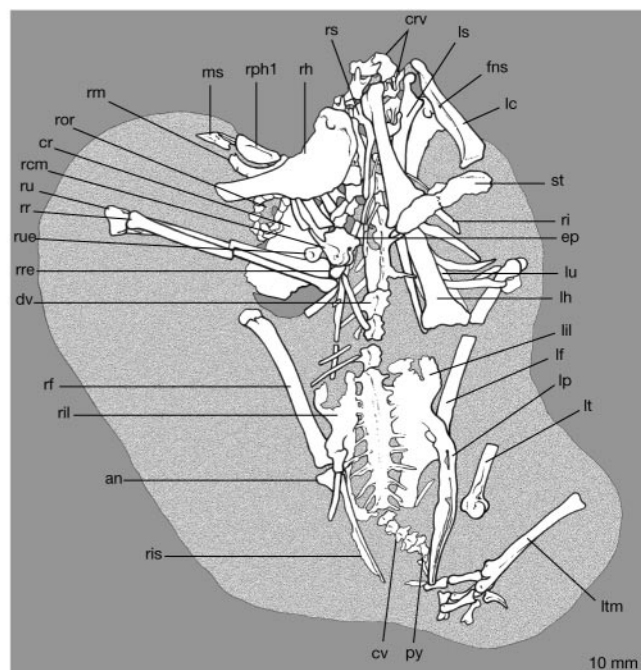
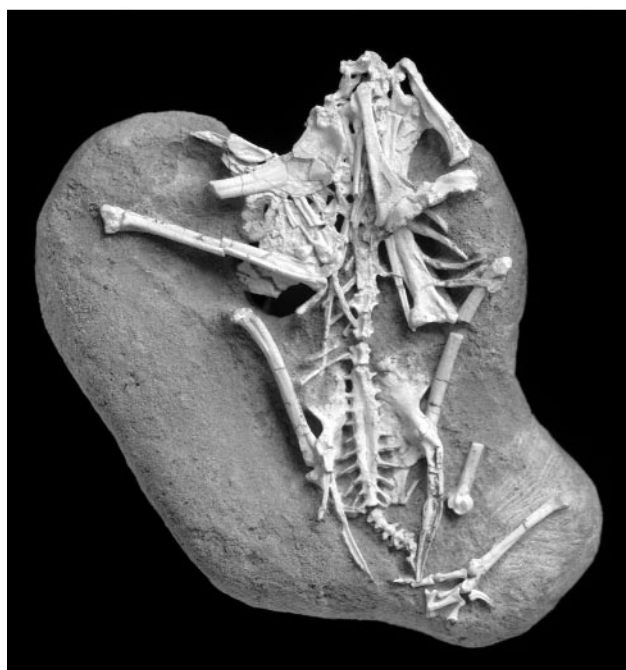


Figure 1 *Apsaravis ukhaana*, holotype IGM 100/1017. Photograph of holotype (left) and annotated diagram (right). an, antitrochanter; c, coracoid; cr, cranium; cv, caudal vertebrae; crv, cervical vertebrae; cm, carpometacarpus; dv, dorsal vertebrae; ep, extensor process; f, femur; fns, foramen n. supracoracoideus; h, humerus; il, ilium; is, ischium; m, mandible; ms, mandibular symphysis; or, orbit; ph1, phalanx 1 manual digit II;

p, pubis; py, pygostyle; r, radius; re, radiale; ri, rib; s, scapula; st, sternum; t, tibiotarsus; tm, tarsometatarsus; u, ulna; ue, ulnare. The 'r' and 'l' prefixes before these abbreviations indicate right and left, respectively. (The distal right humerus was removed for further preparation; see Fig. 3.)

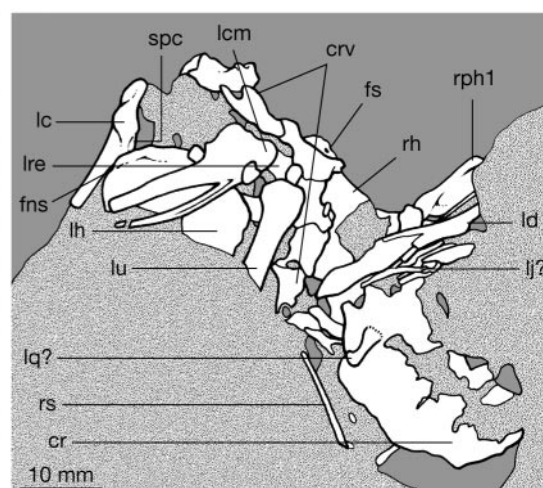


Figure 2 *Apsaravis ukhaana*, holotype IGM 100/1017. Photograph of holotype (left) and annotated diagram (right). c, coracoid; cr, cranium; crv, cervical vertebrae; cm, carpometacarpus; d, dentary; fs, fossa on anterior bicipital crest; fns, foramen n.

supracoracoideus; h, humerus; j?, jugal?; ph1, phalanx 1 manual digit II; q?, quadrate?; re, radiale; s, scapula; spc, scapular cotyla; u, ulna. The 'r' and 'l' prefixes before these abbreviations indicate right and left, respectively.

in *Apsaravis* and Enantiornithes^{11–15}, there is a deep ‘dorsal fossa’ (or a strong concavity) of the posterodorsal coracoid. Because this concavity is present in non-avian theropods¹⁷, as well as in *Confuciusornis*, it is loss of the fossa in Ornithurae that is derived within Avialae.

In *Apsaravis*, the supracoracoideus nerve foramen opens into the proximal ‘dorsal fossa’ (Fig. 2), a condition present within Enantiornithes (such as *Neuquenornis volans*¹⁵). The second, antero-medial, opening of the foramen in Enantiornithes lies in an elongate depression, considered synapomorphic of that clade^{11,12,14}. However, as the foramen lies in this depression in *Apsaravis* (Fig. 1) and the

outgroups¹⁷, this morphology, like the fossa, is also primitive rather than a synapomorphy of Enantiornithes.

A pit-shaped fossa on an anteriorly projected bicipital crest have both been described as synapomorphies of Enantiornithes^{11,12}. But these conditions are also present in *Apsaravis* and in synapomorphies of a more inclusive avialan clade (Fig. 4). Slight anterior projection of the bicipital crest is also known in *Ichthyornis*. The distal humerus (Fig. 3) in *Apsaravis* displays other features considered characteristic of Enantiornithes^{11,13}, including (1) antero-posterior compression and dorsoventral expansion of the distal humerus giving it a ‘spatulate’ aspect¹¹; (2) the angle of the dorsal

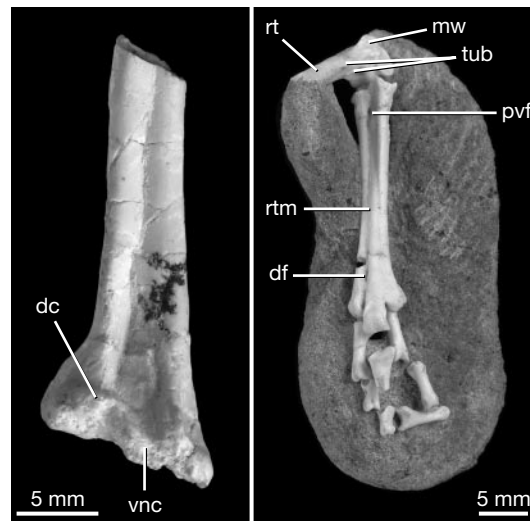


Figure 3 *Apsaravis ukhaana*, holotype IGM 100/1017. Left, left humerus in anterior view; right, right tibiotarsus and tarsometatarsus. dc, dorsal condyle; df, distal foramen; pvf, proximal vascular foramen; t, tibiotarsus; tm, tarsometatarsus; tub, tuberositas retinaculi

extensoris; mw, medial wing of sulcus cartilaginis tibialis; vnc, ventral condyle. The ‘r’ and ‘l’ prefixes before these abbreviations indicate right and left, respectively.

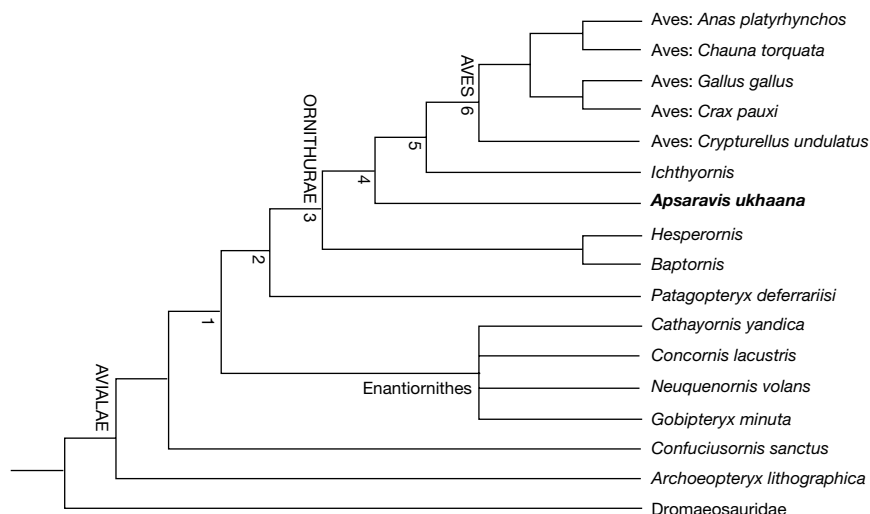


Figure 4 Strict consensus of five most parsimonious cladograms (length, 354; CI, 0.69; RI, 0.81; RC, 0.56) indicating the placement of *Apsaravis ukhaana* in an analysis (using PAUP*4.0b2a²⁵) of 199 characters and 17 taxa (see Supplementary Information). The tree is rooted on Dromaeosauridae (*Velociraptor* and *Deinonychus*). Using *Archaeopteryx lithographica* and *Confuciusornis sanctus* as outgroups does not change the resultant topology. Possible character states at the base of Aves were estimated by the inclusion of taxa sampling clades supported to be basal to Aves (see ref. 6): Anseriformes (for example, *Anas platyrhynchos*, *Chauna torquata*) Galliformes (*Gallus gallus*, *Crax pauxi*) and Palaeognathae (for example, *Crypturellus undulatus*). *Apsaravis* is diagnosed by the following characters (numbers refer to Supplementary Information): dentary strongly forked posteriorly (42); scapular blade narrow throughout its length (102); dorsal humeral condyle angled transversely at high angle to axis of shaft (121); strap-like distal

humeral condyles (122); humeral distal margin angled (123) and anteroposteriorly compressed (124); lateral condyle of tibiotarsus wider than the medial (182); neither condyle of the tibiotarsus medially constricted (183); and the intercondylar groove less than one-third the distal width of the tibiotarsus (184). Placing *Apsaravis* below *Hesperornis* results in only one additional character change: Hesperornithiformes are so apomorphic that few of the 63 included thoracic characters could be scored. The following nodes are supported by the following unambiguous synapomorphies: node 1, 61, 83, 99, 108, 111, 112, 118, 131, 139, 146, 179, 189; node 2, 66, 101, 139, 153, 163, 166, 169, 170, 188, 192; node 3, 30, 31, 53, 54, 61, 156, 157, 165, 172, 178, 182, 190; node 4, 74, 105, 156, 161, 164; node 5, 85, 95, 98, 125, 192; node 6, 8, 37, 55, 61, 90, 96, 113, 114, 115, 118, 119, 133, 147, 149, 179, 180, 193, 201.

condyle relative to long axis of the humerus; and (3) the ventral condyle developed as a weak ridge at the (4) angled distal edge of the humerus.

The *Apsaravis* tibiotarsus has a narrow intercondylar groove¹³ and 'barrel-shaped', medially untapering condyles (Fig. 3). The first character has been considered an enantiornithine synapomorphy^{11–15}. Both features are known from two fragmentary specimens: *Nanantius eos*¹⁸, identified as an enantiornithine; and *Vorona berivotensis*¹⁹, placed in an unresolved trichotomy with Enantiornithes and avialans more closely related to Aves¹⁹. However, *Nanantius eos*, other Enantiornithes, *Vorona* and *Patagopteryx* exhibit the primitive theropod condition^{3,20}, with the medial condyle wider than the lateral. In *Apsaravis* the medial condyle is extremely narrow, whereas most other ornithurines have a broad intercondylar groove and the condyles are roughly equal in width.

To place *Apsaravis* in Enantiornithes requires extensive convergence throughout the skeleton (17+ additional steps). Enantiornithine monophyly, while supported (Fig. 4), is not consistently diagnosed by any single unambiguous synapomorphy in all trees. The distribution of proposed enantiornithine synapomorphies and their status following this analysis is shown in Table 1 of the Supplementary Information. Out of 30 characters considered characteristic of Enantiornithes, only 12 are found exclusively in that clade. Of those that varied across Avialae, 11 apply to more inclusive clades, and the remaining characters either optimize ambiguously or as convergently present in at least one non-enantiornithine clade. Caution is suggested when referring material to Enantiornithes that is based on the simple presence of one or more of these characters in a fragmentary specimen.

This phylogenetic analysis presents further evidence against a monophyletic 'Sauriurae'^{24,21}. The proponents of this proposed clade (including *Archaeopteryx* + *Confuciusornis* + Enantiornithes) invoke convergence to explain characters shared by Enantiornithes and Aves². Many characters used to define separate sauriurine and ornithurine clades²¹ are found in *Apsaravis*. *Apsaravis* displays many characters placing it close to Aves, but retains plesiomorphies present in Enantiornithes and more basal theropods (Fig. 4).

It has been suggested that Enantiornithes were the 'dominant land birds' of the Cretaceous, whereas ornithurines were restricted to near-shore environments and shorebird ecologies². These 'transitional shorebirds' are supposed to have survived a terminal Cretaceous extinction to give rise to the extant radiation of birds in the early Tertiary². This pattern is based on inferring ecological habits from a small number of specimens, which often comprise single bones. *Apsaravis* provides contradictory evidence as it is an ornithurine from continental (not near-shore or marine) deposits and lacks derived affinity with *Ichthyornis* or shorebirds (such as Charadriiformes). Whatever the range of morphologies (and presumed ecology) of Mesozoic ornithurines, there is no evidence that they were stuck in an 'ecological bottleneck'. It may be considered not only a preservational artefact, but also a result of poorly supported referral of fragments to both Enantiornithes and *Ichthyornis* (J. Clarke, manuscript in preparation) (and other taxa considered to have a 'shorebird' ecology¹⁰).

Apsaravis offers insight into the assembly of the modern flight apparatus. *Apsaravis* is the most basal avialan with an extensor process (a strong, anteriorly projected muscular process on metacarpal I) (Fig. 1) related to the insertion of the extensor carpi radialis muscle and the propatagial ligaments in Aves²². In *Archaeopteryx* and non-avian theropods, the proximoanterior edge of metacarpal I flares slightly, but does not extend past the anterior limit of the semilunate carpal. In *Confuciusornis* and Enantiornithes, the entire surface is broadly convex⁶.

In Aves, the extensor carpi radialis muscle is typically the largest forelimb muscle²². During the upstroke–downstroke transition, it automatically extends the manus and attached primary flight feathers to reposition quickly this distal part of the wing for the

downstroke²³. Avian automatic extension of the manus with extension of the forearm was described in non-avian theropods as a consequence of a semilunate trochlear surface in the wrist²⁴. This quick motion, repositioning the manus, was thought to be a primitive action co-opted for flight²⁴.

There is considerable morphological evidence that many aspects of the movements used in powered flight were present before the evolution of flight^{1,9,17}. Theropod dinosaurs more primitive than *Apsaravis* may have had a mechanism enabling automatic extension. However, automatic extension of the manus in Aves is affected primarily by muscles that span both elbow and wrist, and that insert on the extensor process of metacarpal I (ref. 23). There is no morphological evidence of avian coordinated extension, a key part of the avian flight stroke, phylogenetically before *Apsaravis*. The key innovation in the evolution of avian flight is better understood not as the gain of a semilunate carpal in non-flying theropod relatives of birds nor feathers nor flight itself, but as the accrual of morphological novelties both before flight and well after its origin. □

Received 30 June; accepted 23 October 2000.

- Padian, K. & Chiappe, L. M. The early evolution of birds. *Biol. Rev.* **73**, 1–42 (1998).
- Feduccia, A. *The Origin and Evolution of Birds* (Yale Univ. Press, New Haven, 1996).
- Hou, L.-H. *Mesozoic Birds of China* (Taiwan Provincial Feng Huang Ku Bird Park, Nan Tou, 1997).
- Hou, L.-H., Martin, L. D., Zhonghe, Z. & Feduccia, A. Early adaptive radiation of birds: evidence from fossils from northeastern China. *Science* **274**, 1164–1167 (1996).
- Chiappe, L. M., Ji, S.-A., Ji, Q. & Norell, M. A. Anatomy and systematics of the Confuciusornithidae (Aves) from the Mesozoic of Northeastern China. *Am. Mus. Novitates* **242**, 1–89 (1999).
- Clarke, J. A. & Chiappe, L. M. A new carinate bird from the Late Cretaceous of Patagonia (Argentina). *Am. Mus. Novitates* (in the press).
- Dashzeveg, D. et al. Unusual preservation in a new vertebrate assemblage from the Late Cretaceous of Mongolia. *Nature* **374**, 446–449 (1995).
- Chiappe, L. M., Norell, M. A. & Clark, J. M. The skull of a relative of the stem-group bird *Mononykus*. *Nature* **392**, 275–278 (1998).
- Gauthier, J. A. Saurischian monophyly and the origin of birds. *Mem. Cal. Acad. Sci.* **8**, 1–55 (1986).
- Marsh, O. C. *Odontornithes: a Monograph on the Extinct Toothed Birds of North America* (US Geological Exploration of the 40th Parallel, US Government Printing Office, Washington DC, 1880).
- Chiappe, L. M. Late Cretaceous birds of southern South America: anatomy and systematics of Enantiornithes and *Patagopteryx deferrariisi*. *Muenchner Geowissenschaften Abh.* **30**, 203–244 (1996).
- Sanz, J. L., Chiappe, L. M. & Buscalioni, A. D. The osteology of *Concornis lacustris* (Aves: Enantiornithes) from the lower Cretaceous of Spain and a re-examination of its phylogenetic relationships. *Am. Mus. Novitates* **3133**, 1–23 (1995).
- Walker, C. New subclass of birds from the Cretaceous of South America. *Nature* **292**, 51–53 (1981).
- Chiappe, L. M. Cretaceous avian remains from Patagonia shed new light on the early radiation of birds. *Alcheringa* **15**, 333–338 (1991).
- Chiappe, L. M. & Calvo, J. O. *Neuquenornis volans*, a new Enantiornithes (Aves) from the Upper Cretaceous of Patagonia (Argentina). *J. Verteb. Paleontol.* **14**, 230–246 (1994).
- Kurochkin, E. N. A New Enantiornithid of the Mongolian Late Cretaceous, and a General Appraisal of the Infraclass Enantiornithes (Aves) (Special Issue, Palaeontology Institute, Moscow, 1996).
- Ostrom, J. H. The pectoral girdle and forelimb function of *Deinonychus* (Reptilia: Saurischia): a correction. *Postilla* **165**, 1–11 (1974).
- Molnar, R. An enantiornithine bird from the lower Cretaceous of Queensland, Australia. *Nature* **322**, 736–738 (1986).
- Forester, C. A., Chiappe, L. M. & Sampson, S. D. The first Cretaceous bird from Madagascar. *Science* **382**, 532–534 (1996).
- Norell, M. A. & Makovicky, P. Important features of the dromaeosaur skeleton II: information from newly collected specimens of *Velociraptor mongoliensis*. *Am. Mus. Novitates* **3282**, 1–44 (1999).
- Martin, L. D. in *Perspectives in Ornithology* (eds Brush, A. H. & Clark, G. A.) 291–338 (Cambridge Univ. Press, New York, 1983).
- Ostrom, J. H. Some hypothetical stages in the evolution of avian flight. *Smithsonian Contrib. Paleobiol.* **27**, 1–21 (1976).
- Vazquez, R. J. The automating skeletal and muscular mechanisms of the avian wing (Aves). *Zoomorphology* **114**, 59–71 (1994).
- Ostrom, J. H. Wing biomechanics and the origin of bird flight. *N. Jb. Geol. Palaont. Abh.* **195**, 253–266 (1995).
- Swofford, P. L. *PAUP* Phylogenetic analysis using parsimony (*and other methods) Version 4.0b2a* (PPC). (Sinauer Associates, Sunderland, 1998).

Supplementary information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank the Mongolian Academy of Sciences and the joint field parties of the MAS and AMNH. Comments from J. Gauthier, L. Chiappe, B. Creisler, S. Gatesy and P. Makovicky improved the manuscript. M. Ellison provided the figures. A. Davidson prepared the specimen. Support was provided by the Division of Paleontology and the Chapman Memorial Fund (AMNH), the Mercedes Benz Corporation, the National Science Foundation Graduate Fellowship and Yale University.

Correspondence and requests for materials should be addressed to J.C.

Correlated evolution of morphology and vocal signal structure in Darwin's finches

Jeffrey Podos

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

Speciation in many animal taxa is catalysed by the evolutionary diversification of mating signals¹. According to classical theories of speciation, mating signals diversify, in part, as an incidental byproduct of adaptation by natural selection to divergent ecologies^{2,3}, although empirical evidence in support of this hypothesis has been limited^{4–6}. Here I show, in Darwin's finches of the Galápagos Islands, that diversification of beak morphology and body size has shaped patterns of vocal signal evolution, such that birds with large beaks and body sizes have evolved songs with comparatively low rates of syllable repetition and narrow frequency bandwidths. The converse is true for small birds. Patterns of correlated evolution among morphology and song are consistent with the hypothesis that beak morphology constrains vocal evolution, with different beak morphologies differentially limiting a bird's ability to modulate vocal tract configurations during song production. These data illustrate how morphological adaptation may drive signal evolution and reproductive isolation, and furthermore identify a possible cause for rapid speciation in Darwin's finches.

The avian vocal tract, comprised of the trachea, larynx and beak, acts as an acoustic resonance filter during sound production, attenuating harmonic overtones and emphasizing fundamental frequencies produced by the sound source^{7,8}. The vocal tract thus enables birds to produce songs that have a high tonal purity or a 'whistle-like' quality. While singing, songbirds (oscine Passeriformes) actively modify the configuration of their vocal tracts, in a manner closely coordinated with the sound source, so as to maintain the vocal tract's filtering function over a wide range of source frequencies⁷. Studies have shown that vocal tract reconfigurations are achieved largely through rapid changes in beak gape, with increases in beak gape accompanying increases in source frequencies, and vice versa^{9,10}. Beak movements are normally very rapid and precise, and when birds' beaks are either temporarily immobilized or hampered, the tonal purity of songs becomes compromised¹¹. These studies together suggest that limits on the dynamics of vocal tract movements during song production may shape patterns of vocal signal evolution^{12,13}. For example, constraints on vocal tract dynamics in emberizid songbirds probably cause trade-offs between temporal and frequency-based song features¹³, and may set limits on the development and evolution of syllable repetition rate¹⁴.

Vocal performance capacities are predicted to vary as a function of vocal tract morphology, and particularly beak morphology^{12,13}. Songbirds with comparatively large and strong beaks, such as those adapted for crushing hard seeds, should face relatively severe performance constraints on vocal tract dynamics. This is because of an intrinsic trade-off in jaw biomechanics between maximal force and velocity; as jaws become adapted for strength, they will be less able to perform the rapid movements required for the production of certain types of songs. By contrast, songbirds that have evolved smaller beaks, such as those adapted to probe for insects, should suffer less severe constraints on vocal performance. The objective of

this study was to assess, in Darwin's finches, the extent to which adaptive diversification of the beak has shaped the evolution of song features related to vocal performance capacities. Darwin's finches are a particularly appropriate songbird group for this study not only because they express extensive diversity in beak morphology¹⁵ and song¹⁶ (Fig. 1), but also because beak morphology in this group has been shown to adapt precisely and rapidly, through natural selection, to ecological conditions such as food type availability and interspecific competition^{17–19}.

I conducted correlation analyses, at intra- and interspecific scales, between morphological features and a composite measure of the temporal and frequency structure of songs, the 'vocal deviation' (Fig. 2). High vocal deviations are indicative of poor vocal performance, and vice versa (see Methods). Intraspecific analyses were restricted to the medium ground finch, *Geospiza fortis*, for which the largest sample was obtained. *G. fortis* on Santa Cruz Island exhibits unusually high variation in morphological features relative to other Darwin's finches¹⁹, and produces a wide diversity of song types¹⁶. The vocal deviation correlated positively and significantly with all beak measurements and with body mass (Fig. 3; $P < 0.005$ for each correlation), but not with tarsus length ($P = 0.573$) or wing chord length ($P = 0.209$). For interspecific analyses, I calculated correlations between average values of morphological variables and minimal values of vocal deviations (Fig. 4a). These correlations were not tested for statistical significance, however, because of the likelihood of statistical non-independence among species samples²⁰. Correlations were thus re-evaluated in a phylogenetic context (Fig. 4b),

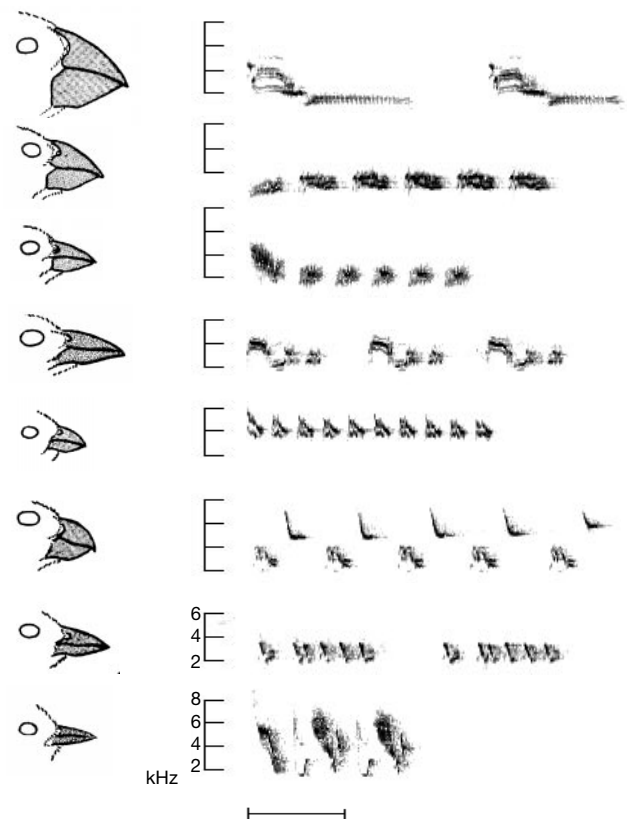


Figure 1 Beak morphology (sketches reprinted¹⁵) and representative sound spectrograms of songs from eight Darwin's finch species on Santa Cruz Island (from top to bottom: *G. magnirostris*, *G. fortis*, *G. fuliginosa*, *G. scandens*, *C. parvulus*, *C. psittacula*, *C. pallida*, *C. olivacea*). Interspecific variation is apparent in both morphology and song structure. Comparability of the songs of different species is supported by the young age of the clade¹⁹, and the striking uniformity among species in the structure of the syrinx and associated musculature²². See ref. 16 for a discussion of homology among Darwin's finch songs. Spectrogram frequency resolution, 98 Hz; scale bar, 0.5 s.

Present address: Department of Biology, University of Massachusetts, Amherst, Massachusetts 01003 USA.

using independent contrasts analyses^{20,21}. Vocal deviation contrasts correlated positively with contrasts of all beak measurements, wing chord length, and body mass (Fig. 4c; $P < 0.051$ for all variables and all phylogenies), but did not correlate with tarsus length contrasts ($P = 0.159$ – 0.207).

The intra- and interspecific analyses are consistent in illustrating that vocal deviations have evolved in correlated fashion with beak morphology, in the direction predicted by the vocal tract constraint

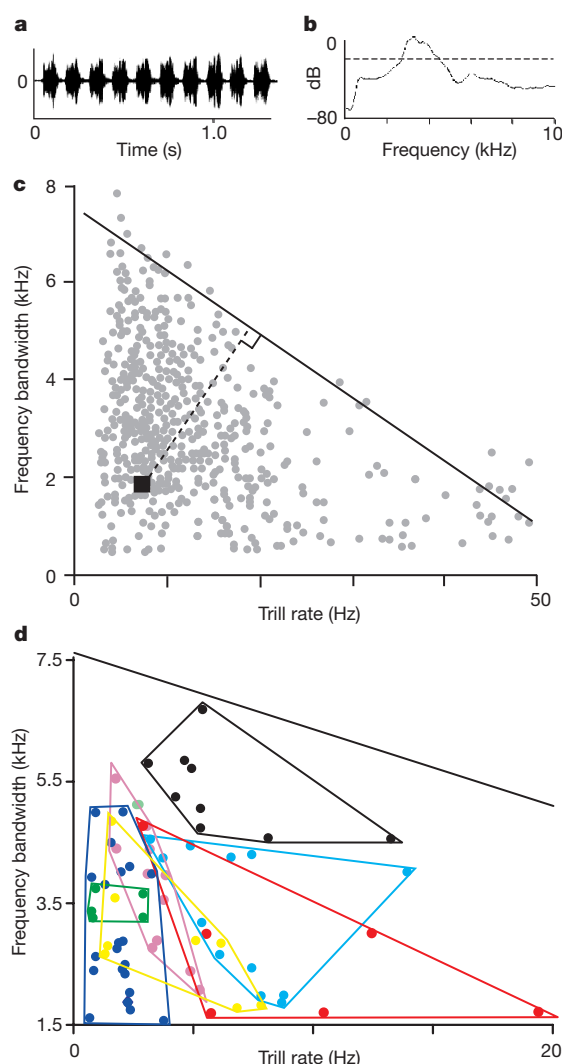


Figure 2 Acoustic analyses. **a–c**, Calculation of trill rate, frequency bandwidth and vocal deviation, illustrated for the *C. parvulus* song in Fig. 1. **a**, Oscillograms were used to measure trill rate, as the number of syllables produced per second (7.85 Hz here). **b**, Amplitude spectra were used to measure frequency bandwidth, the range of frequencies produced (1.97 kHz here), using a -24 dB amplitude cut-off criterion¹³ (dashed line). This criterion was chosen *a priori* to maximize the proportion of signal energy analysed while excluding background noise. Earlier analyses revealed that lower dB cut-off values (for example, -27 and -30 dB) regularly include background noise in spectral analyses, while higher cut-off values (for example, -21 and -18 dB) unnecessarily exclude signal energy. Amplitude spectra were computed at 32 kilopoints and smoothed to a frequency resolution of 300 Hz. **c**, For each trill type, average frequency bandwidth was plotted as a function of average trill rate (for example, filled square). Vocal deviations (dashed line) were calculated relative to an upper-bound regression (solid line)¹³, which was in turn calculated relative to the distribution of trill types across 34 emberizid songbird species (grey dots). **d**, Plot of all trill types for all species analysed, with minimum area convex polygons shown. Overlap in raw data among most species is apparent. *G. magnirostris*, green; *G. fortis*, dark blue; *G. fuliginosa*, pink; *G. scandens*, yellow; *C. parvulus*, light blue; *C. psittacula*, light green; *C. pallida*, red; *C. olivacea*, black.

hypothesis. Vocal deviations also correlated strongly and positively with body mass in both analyses, however, so the potential influence of body mass on vocal performance should be considered as well. Body mass varies positively and nearly isometrically with the size of the syrinx (the sound organ) in Darwin's finches²², and variation in syrinx size presumably influences the expression of fundamental frequency^{16,23}. By contrast there are no clear predictions about how variation in body mass might shape the expression of vocal deviations. As an indirect test of the body mass hypothesis, I assessed the relationship between body mass and minimum vocal deviations across 30 emberizid species¹³, which express a range of body size variation comparable to Darwin's finches. Minimum vocal deviations were found to regress negatively on body mass (slope = -0.028 , $r^2 = 0.038$), in the direction opposite to that predicted by the body mass hypothesis. Although the statistical significance of this regression could not be assessed (because of non-independence among species samples), the negative slope of the regression, along with the finding that body size does not influence emberizid upper-bound regression values¹³, suggest that the vocal tract constraint hypothesis offers a more reasonable explanation for the observed correlations. Body mass in both intra- and interspecific analyses, and wing chord length in the interspecific analyses, might covary with vocal deviations simply because beak and body size measures in Darwin's finches are generally highly correlated¹⁹.

The correlations between morphology (beak and body size) and song offer strong evidence that adaptation can drive signal evolution, as has been discussed in classical theories of speciation but for which little empirical evidence has been offered. Darwin's finches provide one of the premier examples of adaptation in nature, with the most pronounced adaptive diversification centring on beak evolution^{17–19}. Body mass in Darwin's finches also evolves through adaptive processes, both as a direct adaptation to divergent environments and as an indirect correlated effect of beak adaptation¹⁹. Studies of finch adaptation have advanced to the point where it is now possible to predict, with high accuracy, how ecological adaptation (for example, in response to changing food availability) will drive morphological evolution²⁴. The present findings extend this model by illustrating how morphological adaptation in turn drives the evolution of vocal performance capacities. The linkage between morphology and song results in part from the fact that two functional systems—that used for feeding and that used for singing—share a common morphology, the beak.

Morphological adaptation probably shapes finch song evolution in concert with other evolutionary factors, including adaptation to varying acoustic environments¹⁶ and drift caused by copying inaccuracies during song learning (that is, cultural evolution²⁵). The relative influence of these factors varies according to song feature, with, for example, copying inaccuracies exerting their greatest influence on the level of phonology²⁵ and with morphological constraints identified here exerting their greatest influence on higher-order timing and frequency features. A distinguishing feature of morphological influences on vocal evolution is that

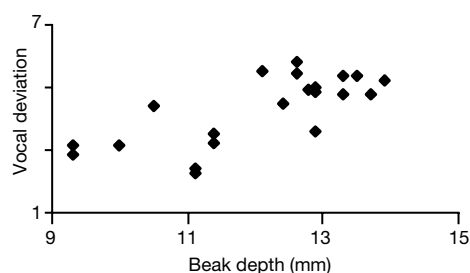


Figure 3 Vocal deviation as a function of beak depth in *G. fortis*. Pearson product moment correlations: beak length, 0.763; beak depth, 0.733; beak width, 0.727; tarsus length, 0.481; wing chord length, 0.549; body mass, 0.792.

they appear to evolve in step with the primary locus of finch adaptation.

Because song presumably has a central role in reproductive isolation in Darwin's finches^{19,26}, the linkage between morphology and vocal performance capacities holds important implications for the dynamics of finch speciation. In island songbird clades, including the Darwin's finches, speciation is driven primarily, if not exclusively, by prezygotic isolating mechanisms, as evident from the regular occurrence of viable and fertile hybrids²⁷. The effectiveness of prezygotic isolating mechanisms generally depends on the extent to which mating signals among incipient species are distinct, with more distinct signals increasing the probability of 'correct' matings and thus enhancing probabilities of speciation⁶. My data suggest that magnitudes of ecological and morphological diversification among incipient Darwin's finch species will directly determine magnitudes of diversification in vocal features, and thus determine probabilities of speciation (to the extent that trill rate and frequency bandwidth are used by birds in mate recognition). Taking this hypothesis one step further, the high diversity of ecological opportunities for Darwin's finches on the Galápagos Islands may thus have promoted, through extensive morphological adaptation and correspondingly large evolutionary changes in vocal signal structure, conditions suitable for rapid speciation and the marked radiation that has defined the group. □

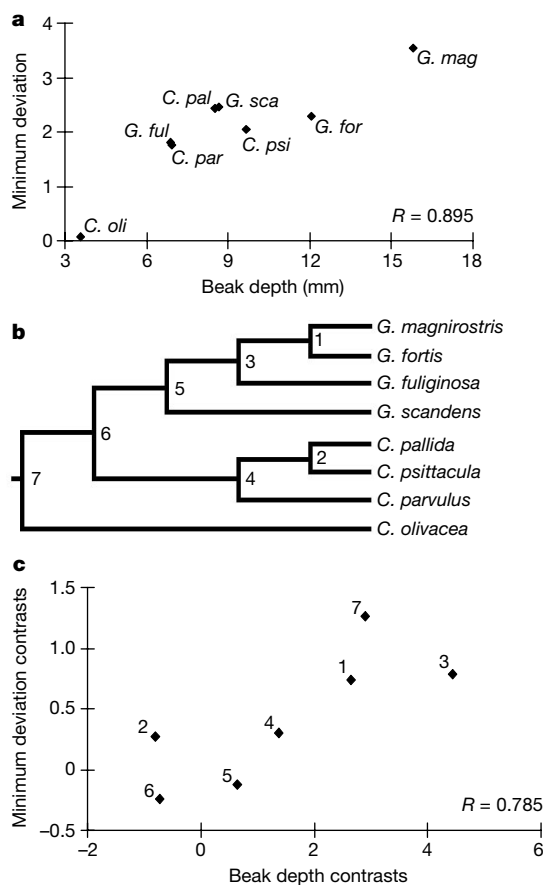


Figure 4 Interspecific analyses. **a**, Minimum vocal deviation as a function of beak depth across eight species of Darwin's finch. **b**, One of four phylogenetic hypotheses used in independent contrasts analyses. Additional hypotheses differ in classifying *G. fortis* as sister taxon to *G. fuliginosa* and/or *C. psittacula* as sister taxon to *C. parvulus*. **c**, Minimum vocal deviation contrasts as a function of beak depth contrasts. Labels refer to phylogenetic nodes from Fig. 4b. Across the four phylogenies, ranges of Pearson product moment correlations were beak length, 0.757–0.809; beak depth, 0.751–0.784; beak width, 0.750–0.794; tarsus length, 0.535–0.588; wing chord length, 0.847–0.869; body mass, 0.835–0.847.

Methods

Sample

I collected morphological and acoustic data from nine species of Darwin's finch. Field work was conducted at coastal and upland sites on Santa Cruz Island during February and March 1999. Male birds, captured in mist nets, were banded with unique colour combinations, measured and released. Morphological measurements included: beak length, beak depth, beak width, tarsus length, wing chord length and body mass¹⁹. Songs of banded birds were recorded using Sennheiser K6/ME66 microphones and Sony TCD-5ProII tape recorders. The number of individuals banded and recorded were as follows: *Geospiza magnirostris*, *n* = 4; *Geospiza fortis*, *n* = 18; *Geospiza fuliginosa*, *n* = 9; *Geospiza scandens*, *n* = 5; *Platyspiza crassirostris*, *n* = 2; *Cactospiza pallida*, *n* = 3; *Camarhynchus psittacula*, *n* = 2; *Camarhynchus parvulus*, *n* = 10; *Certhidea olivacea*, *n* = 7. All song figures were generated and acoustic analyses conducted using SIGNAL version 3.1 sound analysis software (Engineering Design, Belmont MA, 1999).

Acoustic analyses

Eight species in the sample produce songs with trilled sequences (Fig. 1; I define a 'trill' as a series of notes or note groups repeated in succession at a constant tempo). In emberizid songbirds, the family that includes the Darwin's finches, trilled sequences in particular are limited in their timing and frequency structure by constraints on vocal performance¹³. I thus focused on trilled sequences in Darwin's finch songs. For each recorded finch trill, I measured trill rate (rates of syllable repetition; Fig. 2a) and frequency bandwidth (ranges of frequencies produced; Fig. 2b). For each trill type, I then plotted average frequency bandwidth as a function of average trill rate (for example, Fig. 2c, filled square). Across the emberizids, trill rate by frequency bandwidth plots regularly yield triangular distributions at species, generic and family levels (Fig. 2c, grey dots illustrate the pattern at the family level¹³). These triangular distributions presumably occur because of motor trade-offs between rates and magnitudes of vocal tract reconfigurations during trill production¹³. Video analyses of vocal motor activity^{9–11} provide evidence that these triangular distributions also reflect a gradient in vocal performance, with trill types that plot near the origin requiring low levels of vocal activity (for example, only minor vocal tract modulations) and with trill types closer to the upper edge of the distribution requiring higher levels of vocal activity (for example, more vigorous vocal tract modulations¹³).

The realized extreme for trill rate and frequency bandwidth production across the family can be characterized using an upper-bound regression (Fig. 2c, solid line¹³). Minimum distances (Fig. 2c dashed line, referred to herein as 'vocal deviations') between each Darwin's finch trill type and the family-wide upper-bound regression were used as a composite indicator of relative vocal performance, with higher deviation values reflecting lower levels of vocal activity, and vice versa. Use of the emberizid upper-bound regression for determining vocal deviations for Darwin's finches is supported by the fact that Darwin's finches are nested phylogenetically within the emberizids. For interspecific analyses, minimum vocal deviations provided an appropriate indicator of a species' vocal potential, because only a subset of vocal renditions are expected to challenge a species' vocal production capacities. Other possible summary values, including average vocal deviations, include submaximal events and thus less precisely reflect the vocal production capacities of a species¹³.

Phylogenetic analyses

Independent contrast analyses used a punctuational or speciation assumption of evolutionary change, with all branch lengths set to unit length, as has been recommended for clades that have undergone adaptive radiations through the occupation of diverse niches^{5,28}. Phylogenetic hypotheses were based on studies using molecular data and microsatellite DNA variation^{29,30}, which have largely supported earlier hypotheses of branching relations among genera¹⁹. Within the constraints of available information four equally likely species-level phylogenies were identified (see Fig. 4b).

Received 18 July; accepted 23 October 2000.

- West-Eberhard, M. J. Sexual selection, social competition, and speciation. *Quart. Rev. Biol.* **58**, 155–183 (1983).
- Dobzhansky, T. *Genetics and the Origin of Species* 3rd edn (Columbia Univ. Press, New York, 1951).
- Mayr, E. *Animal Species and Evolution* (Harvard Univ. Press, Cambridge, MA, 1963).
- Rice, W. R. & Hostert, E. E. Laboratory experiments on speciation: what have we learned in 40 years? *Evolution* **47**, 1637–1653 (1993).
- Schluter, D. & Nagel, L. Parallel speciation by natural selection. *Am. Nat.* **146**, 292–301 (1995).
- Rundle, H. D., Nagel, L., Boughman, J. W. & Schluter, D. Natural selection and parallel speciation in sympatric sticklebacks. *Science* **287**, 306–308 (2000).
- Nowicki, S. Vocal tract resonances in oscine bird sound production: evidence from birdsongs in a helium atmosphere. *Nature* **325**, 53–55 (1987).
- Fletcher, N. H. & Tarnopolsky, A. Acoustics of the avian vocal tract. *J. Acoust. Soc. Am.* **105**, 35–49 (1999).
- Westneat, M. W., Long, J. H. Jr., Hoese, W. & Nowicki, S. Kinematics of birdsong: functional correlation of cranial movements and acoustic features in sparrows. *J. Exp. Biol.* **182**, 147–171 (1993).
- Podos, J., Sherer, J., Peters, S. & Nowicki, S. Ontogeny of vocal tract movements during song production in the song sparrow. *Anim. Behav.* **50**, 1287–1296 (1995).
- Hoese, W. J., Podos, J., Boetticher, N. C. & Nowicki, S. Vocal tract function in birdsong production: experimental manipulation of beak movements. *J. Exp. Biol.* **203**, 1845–1855 (2000).
- Nowicki, S., Westneat, M. W. & Hoese, W. Birdsong: motor function and the evolution of communication. *Sem. Neurosci.* **4**, 385–390 (1992).
- Podos, J. A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). *Evolution* **51**, 537–551 (1997).
- Podos, J. Motor constraints on vocal development in a songbird. *Anim. Behav.* **51**, 1061–1070 (1996).
- Bowman, R. I. Morphological differentiation and adaptation in the Galápagos finches. *Univ. Calif. Publ. Zool.* **58**, 1–302 (1961).

16. Bowman, R. I. in *Patterns of Evolution in Galápagos Organisms* (eds Bowman, R. I., Berson, M. & Leviton, A. E.) 237–537 (American Association for the Advancement of Science, Pacific Division, San Francisco, 1983).
17. Schluter, D., Price, T. D. & Grant, P. R. Ecological character displacement in Darwin's finches. *Science* **227**, 1056–1059 (1985).
18. Gibbs, H. L. & Grant, P. R. Oscillating selection in Darwin's finches. *Nature* **327**, 511–513 (1987).
19. Grant, P. R. *Ecology and Evolution of Darwin's Finches* 2nd edn (Princeton Univ. Press, Princeton, 1999).
20. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1–25 (1985).
21. Martins, E. P. *COMPARE, version 4.2. Computer Programs for the Statistical Analysis of Comparative Data* (Univ. Oregon, Eugene, Oregon, 1999).
22. Cutler, B. *Anatomical Studies on the Syrinx of Darwin's Finches*. Thesis, San Francisco State Univ. (1970).
23. Ryan, M. J. & Brenowitz, E. A. The role of body size, phylogeny, and ambient noise in the evolution of bird song. *Am. Nat.* **126**, 87–100 (1985).
24. Grant, P. R. & Grant, B. R. Predicting microevolutionary responses to directional selection on heritable variation. *Evolution* **49**, 241–251 (1995).
25. Grant, P. R. & Grant, B. R. Cultural inheritance of song and its role in the evolution of Darwin's finches. *Evolution* **50**, 2471–2487 (1996).
26. Ratcliffe, L. M. & Grant, P. R. Species recognition in Darwin's finches (*Geospiza*, Gould. III. Male responses to playback of different song types, dialects and heterospecific songs. *Anim. Behav.* **33**, 290–307 (1985).
27. Grant, P. R. & Grant, B. R. Speciation and hybridization in island birds. *Phil. Trans. R. Soc. Lond. B* **351**, 765–772 (1996).
28. Mooers, A. Ø., Vamori, S. M. & Schluter, D. Using phylogenies to test macroevolutionary hypotheses of trait evolution in cranes (Gruinae). *Am. Nat.* **154**, 249–259 (1999).
29. Petren, K., Grant, B. R. & Grant, P. R. A phylogeny of Darwin's finches based on microsatellite DNA length variation. *Proc. R. Soc. Lond. Ser. B* **266**, 321–330 (1999).
30. Sato, A. *et al.* Phylogeny of Darwin's finches as revealed by mtDNA sequences. *Proc. Natl Acad. Sci. USA* **96**, 5101–5106 (1999).

Acknowledgements

Field work was coordinated through the Charles Darwin Research Station and the Galápagos National Park Service. I thank M. Rossi-Santos, M. Moreano, H. Vargas, H. Snell and M. Hau for assistance in the field; S. Nowicki, L. Baptista, R. Bensted-Smith, R. Bowman, P. & R. Grant, A. Hendry, W. Hoese, S. Hopp, J. Jaenike, J. Lundberg, W. Maddison, L. McDade, C. Nufio, D. Papaj, S. Patek and R. Prum for discussion and feedback; and the National Science Foundation, the Univ. Arizona Foundation, the Univ. Arizona Office of the Vice President for Research and TAME airlines for financial support.

Correspondence and requests for materials should be addressed to the author (e-mail: jpodos@bio.umass.edu).

Nitrogen limitation of microbial decomposition in a grassland under elevated CO₂

S. Hu^{*,†}, F. S. Chapin III^{*,†}, M. K. Firestone[‡], C. B. Field[§] & N. R. Chiariello[§]

^{*} Department of Integrative Biology, University of California, Berkeley, California 94720, USA

[‡] Department of Environmental Science, Policy and Management, University of California, Berkeley, California 94720, USA

[§] Carnegie Institution of Washington, Department of Plant Biology, Stanford, California 94305, USA

Carbon accumulation in the terrestrial biosphere could partially offset the effects of anthropogenic CO₂ emissions on atmospheric CO₂ (refs 1, 2). The net impact of increased CO₂ on the carbon balance of terrestrial ecosystems is unclear, however, because elevated CO₂ effects on carbon input to soils and plant use of water and nutrients often have contrasting effects on microbial processes^{3–5}. Here we show suppression of microbial decomposition in an annual grassland after continuous exposure to increased CO₂ for five growing seasons. The increased CO₂ enhanced plant nitrogen uptake, microbial biomass carbon, and

available carbon for microbes. But it reduced available soil nitrogen, exacerbated nitrogen constraints on microbes, and reduced microbial respiration per unit biomass. These results indicate that increased CO₂ can alter the interaction between plants and microbes in favour of plant utilization of nitrogen, thereby slowing microbial decomposition and increasing ecosystem carbon accumulation.

Terrestrial ecosystems contain nearly three times more C (~2,060 Gt) than the atmosphere (~735 Gt C)⁶ and may be either a significant C sink or source under future CO₂ models⁷. CO₂ enrichment often enhances ecosystem C gain in the short term through the stimulation of photosynthesis². Over the long term, however, ecosystem C content depends on the balance between net primary production (NPP) and decomposition. Although increased CO₂ affects microbial processes by increasing C inputs to soil^{3,4} and soil moisture^{8,9}, future ecosystem C content has been considered primarily in terms of nutrient supply to plants (C inputs) rather than CO₂ effects on decomposition (C loss). A large source of uncertainty is the effect of increased CO₂ on N availability to plants and microbes¹⁰. Because plants are commonly N-limited in terrestrial ecosystems^{10,11}, any tendency for increased CO₂ to decrease N availability can suppress the stimulation of NPP by elevated CO₂. Microbes can also be N limited^{12,13}, and a decrease in N availability due to increased CO₂ could decrease decomposition and enhance C storage. Conversely, N stimulation of C fixation under increased CO₂ may not necessarily translate into ecosystem C storage because increased N availability can also stimulate microbial decomposition and C turnover¹⁴.

We used a sandstone grassland with moderate soil fertility to investigate the effect of increased CO₂ on microbial biomass and activity and on the interaction between plants and microbes in N acquisition. Experiments were conducted in an annual grassland at Stanford University's Jasper Ridge Biological Preserve in central California (37° 24' N, 122° 14' W; elevation 150 m) between 1992 and 1997. The climate is mediterranean with cool, wet winters and dry summers. Two CO₂ concentrations, ambient (360 p.p.m.) and increased (720 p.p.m.), were maintained for five years with open-top chambers (ten replicates for each treatment)^{15,16}. Soil samples were collected on 26 November 1996, 23 March and 23 April 1997, corresponding approximately to early germination, peak physiological activity, and peak biomass of plants, respectively.

Annual NPP was either stimulated or not affected by increased CO₂ (refs 15, 16). By the end of the sixth growing season, plots exposed to increased CO₂ exhibited a moderate increase (2,750 g C m⁻²) in the stock of below-ground organic C (soil, debris plus roots) compared with ambient controls (2,612 g C m⁻²). Although this increase is not significant at *P* = 0.05, it is consistent with a moderate increase in C inputs¹⁶. We studied organic C

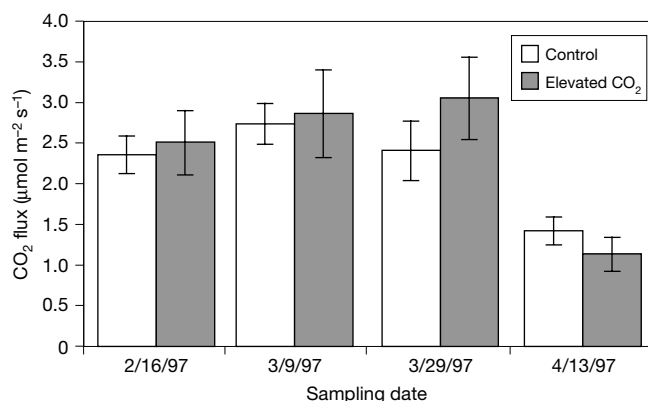


Figure 1 Below-ground respiration. Ambient (open bars) and elevated (filled bars) CO₂. Values are means ± standard error of the mean, s.e.m., *n* = 5–6 plots.

[†] Present address: Department of Plant Pathology, North Carolina State University, Raleigh, North Carolina 27695, USA (S.H.); Institute of Arctic Biology, University of Alaska, Fairbanks, Alaska 99775, USA (F.S.C.).

turnover in the elevated CO_2 plots, based on the lower ^{13}C of the elevated CO_2 . About 22% of total below-ground organic C originated from photosynthetic products during the 5.5-yr experiment. We also examined below-ground respiration (BR, the sum of root (RR) and heterotrophic microbial respiration (HR)) from mid-February to mid-April of 1997, the most active period of plant growth and decomposition. In contrast to the stimulation of BR by elevated CO_2 during the second year of the experiment (the 1993–94 growing season)¹⁷, average BR was not significantly affected by CO_2 ($P = 0.514$; Fig. 1). This low response of BR to increased CO_2 occurred even though microbial biomass increased (see below). The C:N ratio of root biomass remained unchanged (data not shown), so respiration rate per unit root mass was probably unaffected by increased CO_2 (refs 18, 19). We conservatively estimated HR in the field by assuming that RR remained unchanged, although root biomass tended to be higher under elevated CO_2 (ref. 16). These estimates of HR per unit of microbial biomass decreased significantly under elevated CO_2 in the late growing season. However, concurrent laboratory incubation experiments indicated that C availability to soil microbes was significantly higher in soils under increased than under ambient CO_2 . When soils sampled on 23 March were incubated at field moisture in the absence of plants, HR in elevated CO_2 plots ($64.81 \pm 2.90 \mu\text{g CO}_2 \text{ soil d}^{-1}$) was higher than in the ambient controls ($55.62 \pm 3.49 \mu\text{g CO}_2 \text{ g}^{-1} \text{ soil d}^{-1}$) ($P = 0.058$), indicating that field HR was neither moisture- nor carbon-limited. Similarly, when soils sampled on 23 April were incubated at a constant moisture (20% w/w), HR in increased- CO_2 plots ($96.89 \pm 6.99 \mu\text{g CO}_2 \text{ g}^{-1} \text{ soil d}^{-1}$) was higher than their ambient controls ($77.41 \pm 4.39 \mu\text{g CO}_2 \text{ g}^{-1} \text{ soil d}^{-1}$) ($P = 0.015$). Together, these results suggest that microbial decomposition processes were suppressed in the middle to late growing season.

Increased CO_2 significantly decreased soil-extractable N late in the growing season (Fig. 2), when plant growth was most rapid. When plants were present (March and April), microbial biomass C was significantly higher in increased than in ambient CO_2 (Fig. 3a), but microbial biomass N remained unaffected by CO_2 increase (Fig. 3b), leading to higher C:N ratios in the microbial biomass (Fig. 3c). These results suggest that elevated CO_2 caused decreased N-availability to microbial biomass or an increased fungal/bacterial ratio. Early in the growing season before plants were major N consumers, microbial biomass C:N ratio was lower ($P = 0.068$), and there were more active bacteria (data not shown; $n = 5$; $P < 0.01$) in elevated than ambient CO_2 . These data indicate that CO_2 enrichment increased N constraints on microbes at critical times during the year, with the net effect on microbes dependent on whether soil microbes were C- or N-limited.

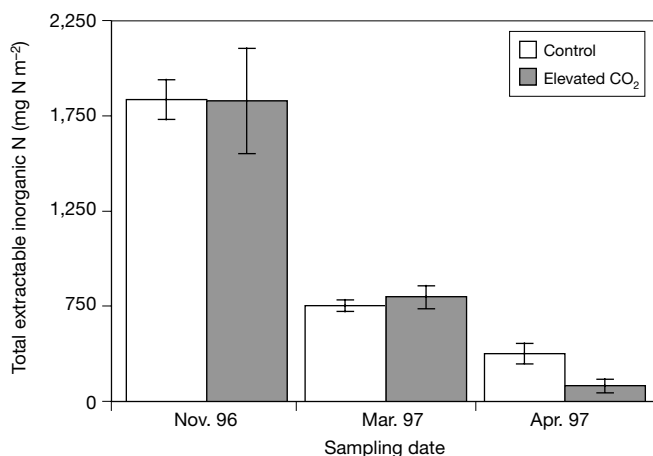


Figure 2 Extractable inorganic soil N over the growing season. Ambient (open bars) and elevated (filled bars) CO_2 . Values are means $\pm$ s.e.m., $n = 10$ plots.

Studies addressing the effects of increased CO_2 on microbes have generally emphasized the importance of enhanced C-input for microbial biomass and activity^{3,4,20}, building on the assumption that microbes in soil are C-limited²¹. Little attention has been directed to the effects of increased CO_2 on plant and microbial use of N. We used ^{15}N as a tracer to investigate the impact of CO_2 enrichment directly on plant–microbial N partitioning during the middle to late growing season when N availability was low (Fig. 2). Results from the ^{15}N partitioning experiment indicate that increased CO_2 increased uptake of N by plants, leading to a substantial increase of ^{15}N in plant shoots and a corresponding decrease in the soil, but no change in the roots ($P = 0.867$) (Fig. 4a). ^{15}N in soil K_2SO_4 extracts and microbial biomass were 31% and 16% higher in ambient than elevated CO_2 , respectively (Fig. 4b). These results show that increased CO_2 altered the interaction between plants and microbes in favour of plant N utilization. In response to N-limitation, plants may increase fine root production and/or mycorrhizal infection may increase²²; both responses would serve to decrease the path length for diffusion of soil N in solution to roots.

Plant responses to increased CO_2 may affect soil microbial biomass and activity, either positively by increasing the availability of C and water, or negatively by reducing litter quality²³. The effect of CO_2 -driven N immobilization by plants on microbes has been largely overlooked. When N is available, increased C input caused by CO_2 enrichment stimulates microbial activity, especially N

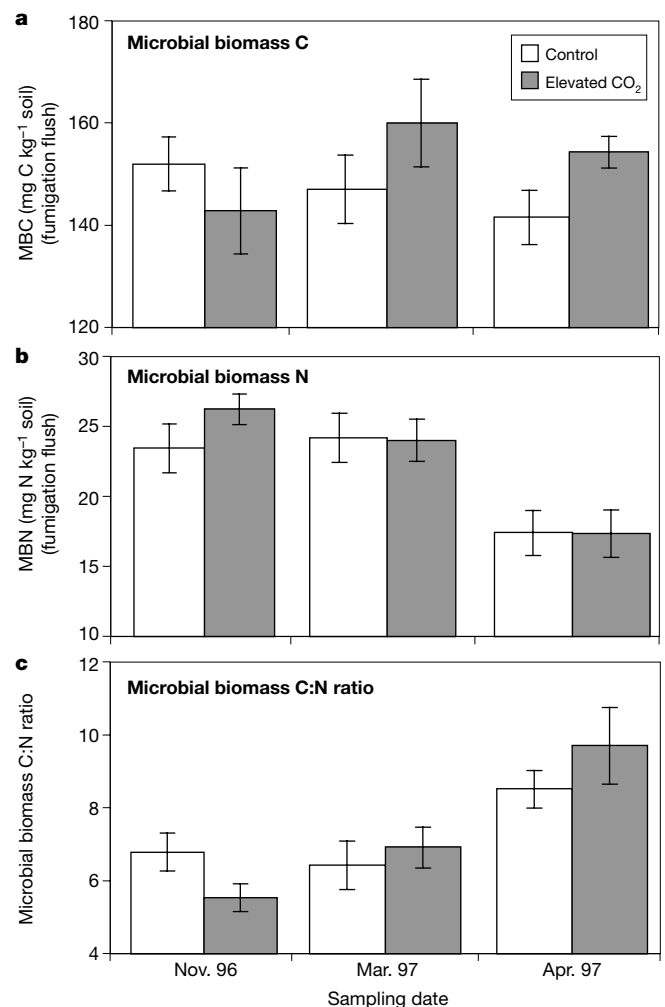


Figure 3 Microbial biomass carbon (a), nitrogen (b) and C:N ratios (c) in ambient (open bars) and elevated (filled bars) CO_2 . Microbial biomass C and N (MBC and MBN) were the flush of organic C or N following the fumigation. Values are means $\pm$ s.e.m., $n = 10$ plots.

utilization^{3,4,14}. Conversely, when N becomes limiting, competition between plants and microbes for N intensifies¹³. Enhanced C input and N limitation under increased CO₂ may favour fungi over bacteria²⁴ because fungal biomass has a lower C:N ratio and fungal hyphae can translocate nutrients. Root infection, hyphal length of arbuscular mycorrhizae, and concentration of glomalin, a glycoprotein produced by arbuscular mycorrhizal fungi, were higher in the elevated CO₂ plots than in the control in this grassland^{25,26}. Enhanced production of fine roots and fungal hyphae could protect soil C by enhancing soil aggregation²⁷. A soil community with greater fungal to bacterial biomass may be conducive to soil C retention^{26,27}. Because the timing of active plant growth largely coincides with active decomposition⁷, CO₂-induced decrease in available N for saprophytic microbes could lead to slower rates of decomposition. Increased plant N uptake under elevated CO₂ may, therefore, lay the foundation for a feedback loop in which the slow decomposition of residues retards the release of nitrogen²⁸, thereby constraining the plant response to increased CO₂.

Our results demonstrate that CO₂-enrichment in the atmosphere can exacerbate N constraints on microbes and suppress microbial use of soil C. Our findings also indicate that CO₂-driven changes in plant-microbe N partitioning could enhance C accumulation in terrestrial ecosystems under future CO₂ levels by increasing plant N-uptake, decreasing N availability for saprophytic microbes, and reducing rates of decomposition. This reduction in decomposition could cause terrestrial ecosystems to become net C sinks, especially if plants become more efficient at acquiring N from low C:N soil organic matter (for example, by increasing mycorrhizae²⁴). Alternatively, if reduced rates of decomposition and resulting soil C accumulation decrease N supply to plants, this could decrease NPP and C inputs to soil, constraining the magnitude of C accumulation.

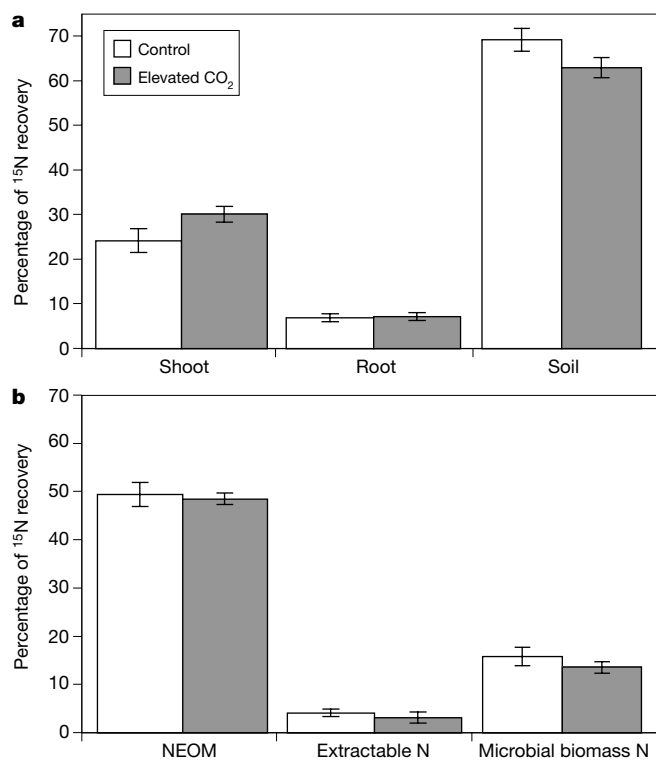


Figure 4 ¹⁵N distribution. **a**, In plant shoots, roots, and soil and **b**, in non-extractable soil organic matter (NEOM), K₂SO₄ extracts and microbial biomass. Ambient (open bars) and elevated (filled bars) CO₂. Values are means $\pm$ s.e.m., $n = 10$ plots. Total microbial biomass ¹⁵N in soil was calculated from microbial biomass ¹⁵N flush by using an extraction efficiency of 0.45.

Depending on the dynamics of N availability and species characteristics, a single ecosystem could experience N limitation predominantly through effects on the microbes at some times and effects on the plants at others. □

Methods

Field treatments and soil sampling

CO₂ fumigation began in January 1992 and has since been continuous except during the summer of 1992 following plant senescence. Open-top chambers were used to maintain the CO₂ concentration at ambient and approximately 720 p.p.m.¹⁵. Two 3.2-cm diameter soil cores were sampled to 15 cm depth from each plot on 26 November 1996 and 23 March 1997, and one 7.2-cm diameter core (15 cm depth) from each plot on 23 April 1997. Soil samples were sieved (2 mm), and roots and debris on the sieve were manually sorted and oven-dried. Soil, shoot, detrital and root carbon and nitrogen, and their isotope ratios were determined by combustion on a gas chromatography-mass spectrometer (GC-MS) (Europa Scientific Ltd.) using finely ground subsamples. We define below-ground organic C as all the organic C except for standing above-ground biomass. Microbial biomass C and N were determined by fumigation-extraction²⁹.

Plant-microbial ¹⁵N partitioning experiment

One aluminium tube (7.2-cm diameter) was installed in each chamber to 20-cm depth on 22 February 1997 and 4.0 mg ¹⁵N of (¹⁵NH₄)₂SO₄, 99.7% atom ¹⁵N) were injected on 8 March 1997 in 25 ml of solution. Above-ground plants and soil cores were collected on 23 April 1997. ¹⁵N content in shoots, roots and soils were determined on a GC-MS as described above. Microbial biomass ¹⁵N ratio was determined after Kjeldahl digestion and diffusion³⁰. Sample $\delta^{15}\text{N}$ (‰) were converted to excess nitrogen isotope (mg); conversion of $\delta^{15}\text{N}$ (‰) to the absolute isotope ratio (¹⁵N/¹⁴N) of the sample was based on the atomic ratio of atmospheric nitrogen. Sample ¹⁵N content was then calculated from fractional abundance (¹⁵N/(¹⁵N + ¹⁴N)) and total N content of the sample. Non-extractable ¹⁵N content (mostly organic ¹⁵N) was calculated by subtracting microbial biomass ¹⁵N and K₂SO₄-extractable ¹⁵N from total soil ¹⁵N.

Root and microbial respiration

Field below-ground respiration (BR) was determined using a LI-COR 6200 CO₂ Analyzer (LI-COR, Inc.)¹⁷. BR was measured in five or six chambers of each treatment with three locations in each chamber. HR in root-free soils was measured as total CO₂ evolution from incubation at 26 °C for 1 day at field moistures or 14 days at constant moisture (20% w/w). The 1-d incubations at field moistures were used to test whether CO₂-improved moistures enhanced HR, whereas the 14-d incubations were to examine the effect of altered C conditions on HR at the constant moisture. CO₂ and C isotope ratios were determined on a GC-MS.

Received 18 October; accepted 3 November 2000.

- Schimel, D. et al. in *Climate Change 1995: The Science of Climate Change* (eds Houghton, J. T. et al.) 65–131 (Cambridge Univ. Press, Cambridge, 1996).
- DeLucia, E. H. et al. Net primary production of a forest ecosystem with experimental CO₂ enrichment. *Science* **284**, 1177–1179 (1999).
- Diaz, S., Grime, J. P., Harris, J. & McPherson, E. Evidence of a feedback mechanism limiting plant response to elevated carbon dioxide. *Nature* **364**, 616–617 (1993).
- Zak, D. R. et al. Elevated atmospheric CO₂ and feedback between carbon and nitrogen cycles. *Plant Soil* **151**, 105–117 (1993).
- Jones, T. H. et al. Impacts of rising atmospheric carbon dioxide on model terrestrial ecosystems. *Science* **280**, 441–443 (1998).
- Schimel, D. S. Terrestrial ecosystems and the carbon cycles. *Glob. Change Biol.* **1**, 77–91 (1995).
- Raich, J. W. & Potter, C. S. Global patterns of carbon dioxide emissions from soils. *Glob. Biogeochem. Cycles* **9**, 23–36 (1995).
- Jackson, R. B., Sala, O. E., Field, C. B. & Mooney, H. A. CO₂ alters water use, carbon gain, and yield for the dominant species in a natural grassland. *Oecologia* **98**, 257–262 (1994).
- Rice, C. W., Garcia, F. O., Hampton, C. O. & Owensby, C. E. Soil microbial response in tallgrass prairie to elevated CO₂. *Plant Soil* **165**, 67–74 (1994).
- Vitousek, P. M. & Howarth, R. W. Nitrogen limitation on land and in the sea—how can it occur. *Biogeochemistry* **13**, 87–115 (1991).
- McGuire, A. D., Melillo, J. M. & Joyce, L. A. The role of nitrogen in the response of forest net primary production to elevated atmospheric carbon dioxide. *Annu. Rev. Ecol. Syst.* **26**, 473–503 (1995).
- Kaye, J. P. & Hart, S. C. Competition for nitrogen between plants and soil microorganisms. *Trends Ecol. Evol.* **12**, 139–143 (1997).
- Wang, J. G. & Bakken, L. R. Competition for nitrogen during mineralization of plant residues in soil: Microbial response to C and N. *Soil Biol. Biochem.* **29**, 163–170 (1997).
- Körner, C. & Arnott, J. A. Responses to elevated carbon dioxide in artificial tropical ecosystems. *Science* **257**, 1672–1675 (1992).
- Field, C. B. et al. in *Carbon Dioxide and Terrestrial Ecosystems* (eds Koch, G. W. & Mooney, H. A.) 121–145 (Academic, San Diego, 1996).
- Hungate, B. A. et al. The fate of carbon in grasslands under carbon dioxide enrichment. *Nature* **388**, 576–579 (1997).
- Luo, Y., Jackson, R. B., Field, C. B. & Mooney, H. A. Elevated CO₂ increases belowground respiration in California grasslands. *Oecologia* **108**, 130–137 (1996).
- Lambers, H., Stulen, I. & Van Der Werf, A. Carbon use in root respiration as affected by elevated atmospheric CO₂. *Plant Soil* **187**, 251–263 (1996).
- Fitter, A. H. et al. Root production and turnover and carbon budgets of two contrasting grasslands under ambient and elevated atmospheric carbon dioxide concentrations. *New Phytol.* **137**, 247–255 (1997).

20. Paterson, E. *et al.* Effect of elevated CO₂ on rhizosphere carbon flow and soil microbial processes. *Glob. Change Biol.* **3**, 363–377 (1997).
21. Smith, J. L. & Paul, E. A. in *Soil Biochemistry* Vol. 6 (eds Bollag, J. & Stotzky, G.) 357–395 (Marcel Dekker, New York, 1990).
22. Rillig, M. C. *et al.* Plant species-specific changes in root inhabiting fungi in a California annual grassland: responses to elevated CO₂ and nutrients. *Oecologia* **113**, 252–259 (1997).
23. Ball, A. S. Microbial decomposition at elevated CO₂ levels: effect of litter quality. *Glob. Change Biol.* **3**, 379–386 (1997).
24. Klironomos, J. N., Rillig, M. C. & Allen, M. F. Below-ground microbial and microfaunal responses to *Artemisia tridentata* grown under elevated atmospheric CO₂. *Funct. Ecol.* **10**, 527–534 (1996).
25. Rillig, M. C., Field, C. B. & Allen, M. F. Soil biota responses to long-term atmospheric CO₂ enrichment in two California annual grasslands. *Oecologia* **119**, 572–577 (1999).
26. Rillig, M. C., Wright, S. F., Allen, M. F. & Field, C. B. Rise in carbon dioxide changes soil structure. *Nature* **400**, 628 (1999).
27. Tisdall, J. M. Possible role of soil microorganisms in aggregation in soils. *Plant Soil* **159**, 115–121 (1994).
28. Bertness, G. M. & Bazzaz, F. A. Belowground positive and negative feedbacks on CO₂ growth enhancement. *Plant Soil* **187**, 119–131 (1996).
29. Vance, E. D., Brookes, P. C. & Jenkinson, D. S. An extraction method for measuring soil microbial biomass C. *Soil Biol. Biochem.* **19**, 703–707 (1987).
30. Stark, J. M. & Hart, S. C. Diffusion technique for preparing salt solutions, Kjeldahl digests, and persulfate digests for nitrogen-15 analysis. *Soil Sci. Soc. Am. J.* **60**, 1846–1855 (1996).

Acknowledgements

We thank H. A. Mooney, C. Lund and B. A. Hungate for contributing to the design and execution of the field experiment, H. L. Zhong for ¹⁵N measurements, and P. Brooks for assistance with ¹³C measurements. The Jasper Ridge CO₂ experiment was supported by grants from the National Science Foundation to the Carnegie Institution of Washington, the University of California, Berkeley, and Stanford University. S.H. was supported by the US NSF under a fellowship awarded in 1996.

Correspondence and requests for materials should be addressed to S.H. (e-mail: shuijin_hu@ncsu.edu).

Regulation of the gain of visually guided smooth-pursuit eye movements by frontal cortex

Masaki Tanaka & Stephen G. Lisberger

Howard Hughes Medical Institute, Department of Physiology, and W. M. Keck Foundation Center for Integrative Neuroscience, University of California, San Francisco, California 94143, USA

In studies of the neural mechanisms giving rise to behaviour, changes in the neural and behavioural responses produced by a given stimulus have been widely reported. This 'gain control' can boost the responses to sensory inputs that are particularly relevant^{1–4}, select among reflexes for execution by motoneurons^{5,6} or emphasize specific movement targets⁷. Gain control is also an integral part of the smooth-pursuit eye movement system^{8–13}. One signature of gain control is that a brief perturbation of a stationary target during fixation causes tiny eye movements, whereas the same perturbation of a moving target during the active state of accurate pursuit causes large responses⁹. Here we show that electrical stimulation of the smooth-pursuit eye movement region in the arcuate sulcus of the frontal lobe ('the frontal pursuit area', FPA) mimics the active state of pursuit. Such stimulation enhances the response to a brief perturbation of target motion, regardless of the direction of motion. We postulate that the FPA sets the gain of pursuit, thereby participating in target selection for pursuit.

We studied sites in the periarculate frontal cortex where electrical stimulation (50 μ A) evoked smooth eye movements when monkeys were performing fixation or pursuit. In agreement with previous reports^{14,15}, these sites were located posterior to and differently from sites where stimulation evoked contraversive saccades. Figure 1a

compares the average eye velocity evoked when a 75-ms train of pulses at 333 Hz was applied during fixation of a stationary target and during pursuit of target motion at 20° s⁻¹ (continuous traces). In each case, the evoked response is compared to the eye velocity recorded in trials with identical target motion but without electrical stimulation (dashed traces). In the example of Fig. 1a, and at almost all our sites, evoked eye movements were larger when stimulation was delivered during pursuit than during fixation. The latency of the responses was slightly but significantly shorter during pursuit than during fixation (mean $\pm$ standard deviation, s.d. is 22.4 $\pm$ 3.2 versus 26.8 $\pm$ 6.2 ms; paired *t*-test, *t* = 5.3, 40 degrees of freedom, d.f., *P* < 0.0001).

The polar plots in Figs 1b and c summarize the direction and amplitude of the eye velocity evoked by electrical stimulation during fixation or pursuit at 20° s⁻¹, for 41 sites in 38 penetrations in two monkeys. The length and orientation of each vector indicate the amplitude and direction of the peak value of evoked eye velocity, respectively. During both fixation and pursuit, the direction of the responses at different sites was widely distributed but showed a strong bias toward ipsiversive (same-sided). The size of the evoked eye movement during pursuit was about three times that during fixation (mean $\pm$ s.d. is 18.1 $\pm$ 8.0 versus 6.2 $\pm$ 4.1° s⁻¹; paired *t*-test, *t* = 14.8, 40 d.f., *P* < 0.0001). However, the direction of the evoked eye movement did not depend strongly on the direction of ongoing pursuit. Regression analysis for the direction of the electrically evoked movement versus the direction of pursuit revealed statistically significant slopes in 10 of 41 sites (*P* < 0.05).

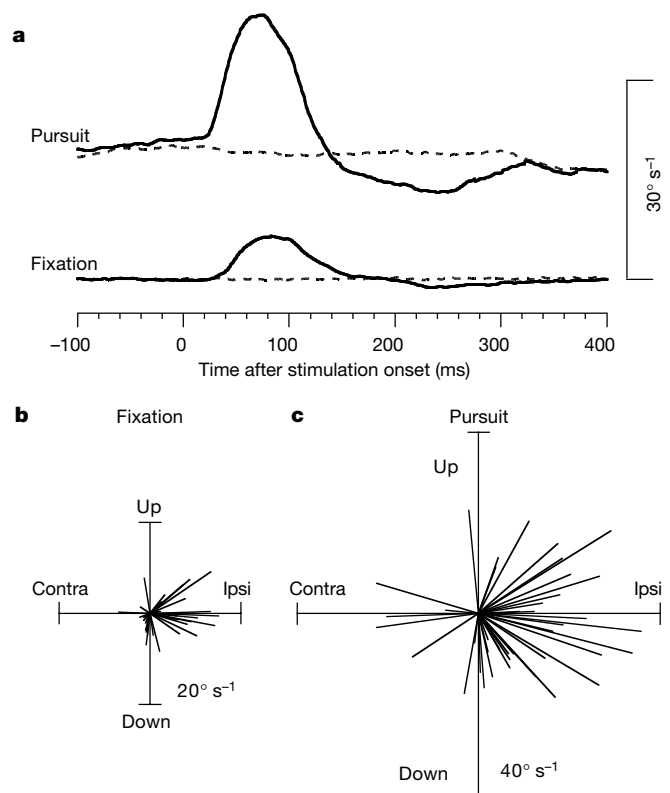


Figure 1 Smooth eye movements evoked by electrical stimulation. **a**, Continuous traces show averages of horizontal eye velocity for trials in which the monkey either fixated a central spot or performed rightward pursuit. Dashed traces show eye movements in non-stimulation controls. In the pursuit task, a stationary target appeared 8° from the centre of the screen for 1,000 to 1,500 ms, then executed step-ramp motion (4° step, 20° s⁻¹ ramp) toward the centre of the screen. Polar plots summarizing the direction and amplitude of elicited eye velocity during fixation (**b**) or pursuit (**c**) at all 41 sites tested. For pursuit trials (**c**), electrical stimulation was applied during pursuit in eight directions, and the largest response is shown.

Only three sites had slopes greater than 0.35 and the mean slope for the remaining 38 sites was 0.11. A slope of 1 would indicate that the direction of elicited movements was the same as that of ongoing pursuit.

We next asked whether electrical stimulation in the FPA could enhance the gain of visual-motor transmission for pursuit, in addition to evoking smooth eye movements. We probed the gain of the pursuit system by presenting a brief perturbation of a stationary target that consisted of a single cycle of a 10-Hz sine wave with an amplitude of $\pm 10^\circ \text{ s}^{-1}$. To ensure conjunction of the electrical stimulation and the 100-ms duration target motion, we delivered trains of electrical stimulation for 200 ms starting at the same time as the perturbation. As the electrical stimulation itself evoked smooth eye movements that would contaminate the image motion provided by the perturbation, we used stimulation with lower-frequency trains of pulses to minimize the velocity of the electrically evoked smooth eye movements. Trains were 100 Hz for 11 sites where stimulation at 333 Hz evoked the largest smooth eye movements during fixation, and 200 Hz for the other 22 sites.

Figure 2a and b illustrates the time course of eye movements obtained from the same stimulation site as in Fig. 1a. The control perturbations presented during fixation in the absence of electrical stimulation caused a small response (dotted traces, labelled 'pert'). The electrical stimulation alone at this site with the longer (200 ms),

lower-frequency (100 Hz) trains caused smooth eye movements with peak velocity of 2.6° s^{-1} when the monkey was fixating a stationary target (dashed traces, labelled 'stim'). Target perturbations during the same electrical stimulation evoked a response (solid traces, labelled 'comb') that was greater than the response to electrical stimulation alone. The component of the combined response driven by target motion was estimated by computing the millisecond-by-millisecond difference: eye velocity evoked by the perturbation in the presence of electrical stimulation minus that evoked by electrical stimulation alone. Comparison of the difference in eye velocity (continuous traces, labelled 'comb - stim') with the responses to the control perturbation in the absence of electrical stimulation (dotted traces, labelled 'pert') reveals that electrical stimulation enhanced the response considerably.

The enhancement was omni-directional: the initial component of the response, in the direction of the initial component of target motion, was enhanced whether the initial phase of the perturbation first presented ipsiversive (Fig. 2a) or contraversive (opposite-sided; Fig. 2b) motion. Each experiment provided multiple repetitions of eight different perturbations with initial components that took the target left, down, right, up, or in one of the four oblique directions. We quantified the response to the initial half-cycle of target perturbation by measuring the peak of the difference in eye velocity during the interval from 100 to 150 ms after the onset of target motion. For the site summarized by the polar plot in Fig. 2c, the responses to perturbations of a stationary target were enhanced during the FPA stimulation (open symbols, labelled 'comb - stim') relative to control (filled symbols, labelled 'pert') for all perturbation directions. In this example, the response is enhanced in the direction of the perturbation.

We performed two statistical tests to verify that the enhancement was omni-directional at most sites. Rayleigh's test for directional uniformity revealed statistically significant ($P < 0.05$) directional bias of the enhancement at only three sites, each of which had tiny enhancements. To test for a directional bias, we used Mardia's circular-linear correlation analysis, which revealed that the magnitude and direction of the changes in responses were not correlated for most sites (27 of 33 sites, $P > 0.05$). Our finding that the response is enhanced in the direction of the perturbation argues that electrical stimulation is enhancing the response to target motion and not vice versa.

To test for directional bias in the population response, we assessed the distribution of the vectors describing the mean enhancement at each site, sorted (1) by the side of the stimulated hemisphere and (2) by the direction of the stimulation-evoked eye movements during fixation at each site. For each case, the distribution was statistically uniform (Rayleigh's test, $P > 0.05$). For the two methods of sorting the data, the mean population vectors were rotated about 45° down relative to (1) ipsiversive and (2) the direction of the evoked eye movement during fixation at each site.

We quantified the 'gain' of the response to perturbations for each condition as described in the Methods. In the example shown in Fig. 2c, electrical stimulation enhanced the gain from 0.11 to 0.33. For comparison with a more natural form of gain enhancement, we recorded the responses to the same target perturbation applied during the maintenance of pursuit, repeating a variation of an experiment reported previously⁹. The data in Fig. 2d were obtained by moving targets in one of eight directions and presenting perturbations along the axis of target motion with each of the two possible polarities. The gain of the response to perturbations was 0.14, 0.24 and 0.32 for target motion of 0, 10 and 30° s^{-1} , respectively. Thus, the electrical stimulation in Fig. 2c caused enhancement that mimicked pursuit of target motion at 30° s^{-1} in Fig. 2d.

Figure 3a summarizes the gain of the responses to perturbation of a stationary target. Enhancement was clear in both monkeys and statistically significant (paired *t*-test, $t = 10.8$, 32 d.f., $P < 0.0001$),

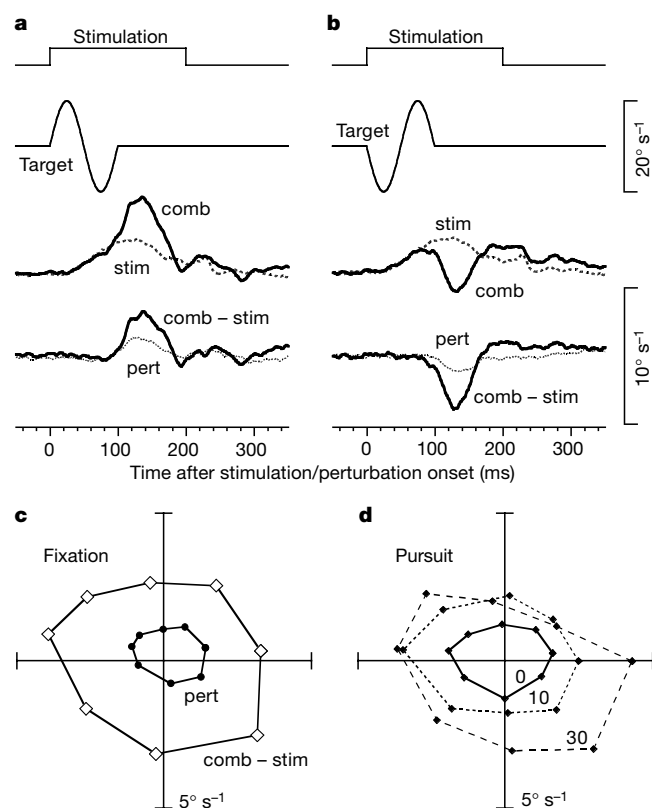


Figure 2 Enhancement of the response to target motion caused by stimulation in the FPA. **a, b**, Different traces show the responses to target perturbation ('pert'), to electrical stimulation ('stim'), and to perturbations during electrical stimulation ('comb'). Traces labelled 'comb - stim' show the difference between the response to the perturbation with electrical stimulation and the response to electrical stimulation alone. **c**, Polar plot showing the responses to perturbations in eight different directions. **d**, Polar plot showing the responses to perturbations delivered during fixation or pursuit in eight directions, without electrical stimulation. Numbers next to the curves indicate target speed. Each point plots the peak of the response to the perturbation in the interval from 100–150 ms after the onset of perturbation. For each direction of perturbation, the responses evoked during opposite pursuit directions were averaged.

despite individual differences in control performance on perturbations in the two monkeys: the one shown with filled symbols had smaller responses to perturbations in the absence of electrical stimulation than did the one shown with open symbols. The ratio of the gains with and without electrical stimulation ranged from 0.93 to 5.63 with a mean of 2.8 (s.d. is 1.1).

Can the enhancement of the response to target motion be accounted for simply as a secondary effect of the smooth eye movements evoked by stimulation itself? In general, the peak eye velocities evoked by electrical stimulation during fixation (abscissa, Fig. 3b) were less than 5° s^{-1} (mean $\pm$ s.d. is $2.8 \pm 1.2^\circ \text{ s}^{-1}$). If the evoked smooth eye movement itself caused the enhancement, then the same amount of enhancement should be present during pursuit at 5° s^{-1} . The leftward arrowheads in Fig. 3b show that this prediction was not fulfilled: they plot the gain of the responses during pursuit at 5° s^{-1} . For one monkey (open symbols), the responses to perturbations delivered during electrical stimulation (ordinate, Fig. 3b) were consistently larger than those delivered during pursuit. For the other monkey, the response during electrical stimulation was as large as or larger than during pursuit at 6 of 8 stimulation sites. Our methods probably underestimate the size of the response to the perturbation during electrical stimulation. Stimulation causes eye velocity, which creates baseline image motion during the perturbation. Other experiments have shown that baseline image motion decreases the gain of the response to a perturbation (M. M. Churchland and S. G. Lisberger, unpublished data).

Stimulating the FPA electrically has two qualitatively different effects. One, expected from previous studies, is injection of a directional eye velocity signal. We have added to previous knowledge by showing that the directional signal is enhanced by the performance of pursuit. This effect seems similar to the results of stimulation experiments done in the middle temporal area^{16,17}, the medial superior temporal area¹⁷, the dorsolateral pontine nucleus¹⁸ and the vermis of the cerebellum¹⁹. The other is an omni-directional enhancement of the response to a brief target perturbation during fixation. This effect seems to set the gain of pursuit, presumably by accessing the same sites that cause the response to target perturbation to be larger during pursuit than during fixation.

Our experiments suggest an unusual view of the role of the FPA. Until now, the FPA has been considered to be a part of the motor

pathway, providing a representation of pursuit direction and/or speed that is passed to the brainstem and the cerebellum. Our data imply that the FPA also plays a role in setting the gain of pursuit. This hypothesis is compatible with the prior studies showing that the FPA contains cells discharging during pursuit^{14,20–22}, and that lesions made in this area reduce the gain of pursuit^{14,23,24}. The dual effect of the FPA stimulation is also compatible with recent suggestions that there are two different classes of signal flow in the pursuit system. One set of signals transforms information about visual motion into commands for smooth eye velocity, while the other set controls the gain of this transformation^{8–12}.

Integration of our findings with other recent reports suggests a relationship between our data and mechanisms of target selection for pursuit. Other studies have demonstrated a post-saccadic enhancement of pursuit that activates the same gain control that we have accessed with stimulation of the FPA¹². If two moving targets are present, then a saccade to one target increases the gain of visual-motor transmission for the motion of the target at the endpoint of the saccade, whatever its direction of motion¹³. Thus, post-saccadic enhancement appears to be part of the causal linkage for selecting targets for pursuit, and acts on the basis of the spatial location of the target rather than its direction of motion. Data from other laboratories also suggest that target selection for pursuit is based on the spatial location of the target rather than its form²⁵ or motion²⁶, implying that the neural substrate of target selection is omni-directional. As the enhancement of the response to a perturbation during stimulation of the FPA is omni-directional, we suggest that the stimulation may access sites, either locally or downstream, that are responsible for target selection for pursuit. Thus, the present study provides a specific example for the broader view that the frontal cortex is involved in voluntary functions such as decision making, or selection of targets and motor programmes^{27–29}. □

Methods

Two male rhesus monkeys were used in experiments that had been approved in advance by the Animal Care and Use Committee of the University of California, San Francisco. The procedures for the surgery and initial behavioural training were the same as those described previously²⁹. During experiments, the monkeys sat in a primate chair with their heads fixed by an implanted head holder. They were required to fixate and track a white target spot (0.2°) that appeared on an analogue oscilloscope (Hewlett-Packard) with a refresh rate of 250 Hz. Eye position was monitored using scleral search coils. Under general anaesthesia and using sterile procedure, a stainless recording cylinder was implanted over the right arcuate sulcus of each monkey.

Stimulation current (cathodal 0.2-ms pulses) was delivered through tungsten electrodes (Frederick Haer & Co.). The intensity of current was $50 \mu\text{A}$, and was monitored by measuring the voltage across a $1 \text{ K}\Omega$ resistor placed in series with the electrode. Sites of stimulation were determined by recording unit activity and by testing the effects of stimulation for every half millimetre along each penetration. Trials with and without electrical stimulation were randomly interleaved. In the first set of experiments, stimulation (75 ms, 333 Hz) was delivered during randomly interleaved fixation trials and pursuit trials. In the pursuit trials, the stimulation was applied when the moving target crossed the centre of the screen, 600 ms after motion onset. The direction of target motion was one of eight radial directions (45° apart) and the target speed was 20° s^{-1} . In the second set of experiments, stimulation (200 ms, 100 or 200 Hz) was applied at the onset of a brief perturbation of a stationary target.

As a control, we tested the responses to the same perturbation during pursuit in the absence of electrical stimulation. The perturbation was introduced 500 ms after the onset of target motion. The direction of perturbation was aligned with the axis of direction of the ongoing target motion, which was again one of eight directions. The initial half-cycle of the sinusoidal perturbation either increased or decreased the speed of the target.

Eye velocity was obtained by analogue differentiation. Data were digitized and sampled at 1 kHz. Saccades seldom occurred during perturbations or the response to perturbations, and were removed from the eye velocity traces. Traces were aligned on the onset of target motion for identical tasks, and the averages of eye velocity were obtained as a function of time. To isolate the response to perturbations, we computed the difference in eye velocity: average eye velocity in trials that included the target perturbation minus eye velocity in trials that were identical except for the absence of a target perturbation.

The onset of the electrically evoked smooth eye movements was detected using a modified technique³⁰. Briefly, baseline eye speed was obtained by computing the mean and standard deviation of angular eye velocity for the 100-ms interval before the stimulation onset. Then a regression line was fitted to the average eye speed for the period from 5 ms before to 10 ms after the average eye speed exceeded the mean ± 2 s.d.. The point where the regression line intersected the baseline mean was taken as the onset of eye movement. We

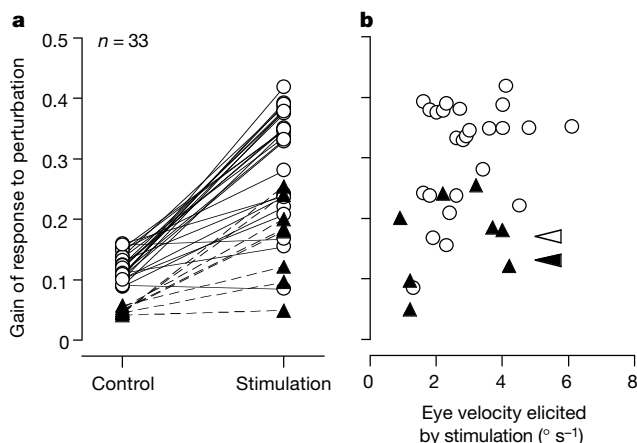


Figure 3 Summary of the enhancement for 33 stimulation sites in 32 penetrations. **a**, Gain of the responses is plotted for different stimulation conditions. Each connected pair of symbols shows data obtained from one experiment. Different symbols are used for different monkeys. **b**, The gain of the response to the perturbation during electrical stimulation is plotted as a function of the peak eye velocity elicited by electrical stimulation in the absence of a perturbation. The two leftward arrowheads plot the gain of the responses to perturbations delivered during pursuit at 5° s^{-1} . Open and filled arrowheads correspond to data plotted as open and filled symbols, respectively.

were unable to use this method for one of 41 sites, since the responses were too small. For this site, the onset of eye movements was detected visually.

For each condition and each stimulation site, the gain of the responses to the perturbations was computed as the square root of E divided by T , where E is the area within the polygon defined by the polar plot for each set of eight perturbations in different directions, and T is the area within the polygon defined by the peak target velocity, always 282.84.

Received 10 July; accepted 17 October 2000.

- Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
- Maunsell, J. H. The brain's visual world: representation of visual targets in cerebral cortex. *Science* **270**, 764–769 (1995).
- True, S. & Martinez Trujillo, J. C. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575–579 (1999).
- Posner, M. I. Attention: the mechanisms of consciousness. *Proc. Natl Acad. Sci. USA* **91**, 7398–7403 (1994).
- Pearson, K. G. Proprioceptive regulation of locomotion. *Curr. Opin. Neurobiol.* **5**, 786–791 (1995).
- Roy, J. E. & Cullen, K. E. A neural correlate for vestibulo-ocular reflex suppression during voluntary eye-head gaze shifts. *Nature Neurosci.* **1**, 404–410 (1998).
- Wurtz, R. H. & Mohler, C. W. Organization of monkey superior colliculus: enhanced visual response of superficial layer cells. *J. Neurophysiol.* **39**, 745–765 (1976).
- Grasse, K. L. & Lisberger, S. G. Analysis of a naturally occurring asymmetry in vertical smooth pursuit eye movements in a monkey. *J. Neurophysiol.* **67**, 164–179 (1992).
- Schwartz, J. D. & Lisberger, S. G. Initial tracking conditions modulate the gain of visuo-motor transmission for smooth pursuit eye movements in monkeys. *Vis. Neurosci.* **11**, 411–424 (1994).
- Krauzlis, R. J. & Miles, F. A. Transitions between pursuit eye movements and fixation in the monkey: dependence on context. *J. Neurophysiol.* **76**, 1622–1638 (1996).
- Keating, E. G. & Pierre, A. Architecture of a gain controller in the pursuit system. *Behav. Brain Res.* **81**, 173–181 (1996).
- Lisberger, S. G. Postsaccadic enhancement of initiation of smooth pursuit eye movements in monkeys. *J. Neurophysiol.* **79**, 1918–1930 (1998).
- Gardner, J. & Lisberger, S. G. Linked target selection for saccadic and smooth pursuit eye movements. *Soc. Neurosci. Abstr.* **25**, 1398 (1999).
- MacAvoy, M. G., Gottlieb, J. P. & Bruce, C. J. Smooth-pursuit eye movement representation in the primate frontal eye field. *Cereb. Cortex* **1**, 95–102 (1991).
- Gottlieb, J. P., Bruce, C. J. & MacAvoy, M. G. Smooth eye movements elicited by microstimulation in the primate frontal eye field. *J. Neurophysiol.* **69**, 786–799 (1993).
- Groh, J. M., Born, R. T. & Newsome, W. T. How is a sensory map read out? Effects of microstimulation in visual area MT on saccades and smooth pursuit eye movements. *J. Neurosci.* **17**, 4312–4330 (1997).
- Komatsu, H. & Wurtz, R. H. Modulation of pursuit eye movements by stimulation of cortical areas MT and MST. *J. Neurophysiol.* **62**, 31–47 (1989).
- May, J. G., Keller, E. L. & Crandall, W. F. Changes in eye velocity during smooth pursuit tracking induced by microstimulation in the dorsolateral pontine nucleus of the macaque. *Soc. Neurosci. Abstr.* **11**, 79 (1985).
- Krauzlis, R. J. & Miles, F. A. Role of the oculomotor vermis in generating pursuit and saccades: effects of microstimulation. *J. Neurophysiol.* **80**, 2046–2062 (1998).
- Gottlieb, J. P., MacAvoy, M. G. & Bruce, C. J. Neural responses related to smooth-pursuit eye movements and their correspondence with electrically elicited smooth eye movements in the primate frontal eye field. *J. Neurophysiol.* **72**, 1634–1653 (1994).
- Tanaka, M. & Fukushima, K. Neuronal responses related to smooth pursuit eye movements in the periaruate cortical area of monkeys. *J. Neurophysiol.* **80**, 28–47 (1998).
- Fukushima, K., Sato, T., Fukushima, Y., Shinmei, Y. & Kaneko, C. R. Activity of smooth pursuit-related neurons in the monkey periaruate cortex during pursuit and passive whole-body rotation. *J. Neurophysiol.* **83**, 563–587 (2000).
- Keating, E. G. Frontal eye field lesions impair predictive and visually-guided pursuit eye movements. *Exp. Brain Res.* **86**, 311–323 (1991).
- Shi, D., Friedman, H. R. & Bruce, C. J. Deficits in smooth-pursuit eye movements after muscimol inactivation within the primate's frontal eye field. *J. Neurophysiol.* **80**, 458–464 (1998).
- Krauzlis, R. J., Zivotofsky, A. Z. & Miles, F. A. Target selection for pursuit and saccadic eye movements in humans. *J. Cogn. Neurosci.* **11**, 641–649 (1999).
- Adler, S. A., Bala, J. K. & Krauzlis, R. J. Effects of prior sensory and motor information on the initiation of pursuit and saccades. *Soc. Neurosci. Abstr.* **26**, 1716 (2000).
- Goldberg, M. E. & Segraves, M. A. in *The Neurobiology of Saccadic Eye Movements, Reviews of Oculomotor Research* Vol. III (eds Wurtz, R. H. & Goldberg, M. E.) 283–313 (Elsevier, Amsterdam, 1989).
- Schall, J. D. & Thompson, K. G. Neural selection and control of visually guided eye movements. *Annu. Rev. Neurosci.* **22**, 241–259 (1999).
- Lisberger, S. G. & Westbrook, L. E. Properties of visual inputs that initiate horizontal smooth pursuit eye movements in monkeys. *J. Neurosci.* **5**, 1662–1673 (1985).
- Carl, J. R. & Gellman, R. S. Human smooth pursuit: stimulus-dependent responses. *J. Neurophysiol.* **57**, 1446–1463 (1987).

Acknowledgements

We thank J. Maunsell and A. Doupe for comments on the manuscript; S. Tokiyama for technical assistance; K. MacLeod and L. Montgomery for surgical assistance; M. Meneses for animal care; S. Ruffner for computer programs; D. Kleinhesselink for network management; K. McGary for electronic devices; and L. Boskai for machinery. This work was supported by HHMI and by a NIH grant to S.G.L.

Correspondence and requests for materials should be addressed to M.T. (e-mail: masaki@phy.ucsf.edu).

A role for ghrelin in the central regulation of feeding

Masamitsu Nakazato*, Noboru Murakami†, Yukari Date*, Masayasu Kojima‡, Hisayuki Matsuo‡, Kenji Kangawa‡ & Shigeru Matsukura*

* Third Department of Internal Medicine, Miyazaki Medical College, Kiyotake, Miyazaki 889-1692, Japan

† Department of Veterinary Physiology, Miyazaki University, Miyazaki 889-2192, Japan

‡ Department of Biochemistry, National Cardiovascular Center Research Institute, Osaka 565-8565, Japan

Ghrelin is an acylated peptide that stimulates the release of growth hormone from the pituitary¹. Ghrelin-producing neurons are located in the hypothalamus, whereas ghrelin receptors are expressed in various regions of the brain^{2–4}, which is indicative of central—and as yet undefined—physiological functions. Here we show that ghrelin is involved in the hypothalamic regulation of energy homeostasis. Intracerebroventricular injections of ghrelin strongly stimulated feeding in rats and increased body weight gain. Ghrelin also increased feeding in rats that are genetically deficient in growth hormone. Anti-ghrelin immunoglobulin G robustly suppressed feeding. After intracerebroventricular ghrelin administration, Fos protein, a marker of neuronal activation⁵, was found in regions of primary importance in the regulation of feeding, including neuropeptide Y⁶ (NPY) neurons and agouti-related protein⁷ (AGRP) neurons. Antibodies and antagonists of NPY and AGRP abolished ghrelin-induced feeding. Ghrelin augmented NPY gene expression and blocked leptin-induced⁸ feeding

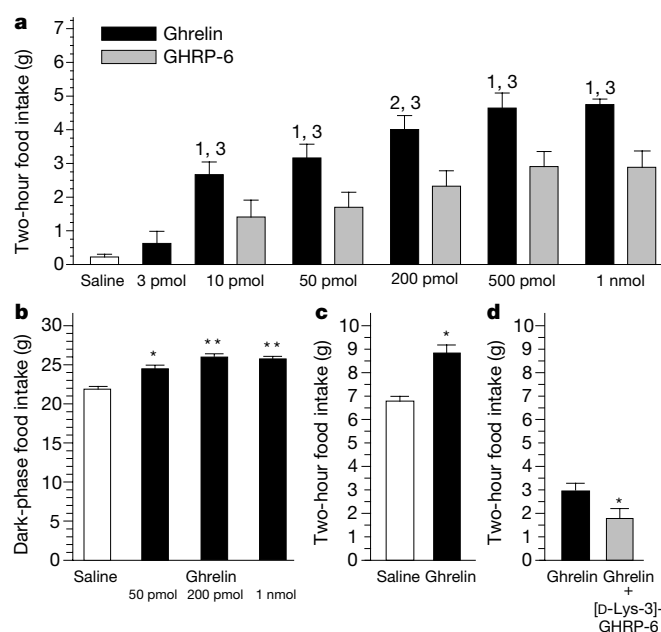


Figure 1 Stimulation of feeding by single ICV administration of ghrelin. **a**, Two-hour food intake (mean $\pm$ s.e.m.) of free-feeding rats injected with various doses of ghrelin or GHRP-6. Control rats were given 0.9% saline. ANOVA was only performed on the ghrelin group against the control group. 1, $P < 0.05$ versus GHRP-6; 2, $P < 0.01$ versus GHRP-6; 3, $P < 0.0001$ versus saline. **b**, Dark-phase (19:00–07:00) food intake of rats receiving an ICV administration of ghrelin at 18:45. Asterisk, $P < 0.005$; double asterisk, $P < 0.001$. **c**, 2-h food intake of 8-h fasted rats receiving ghrelin (200 pmol) at 08:45. Asterisk, $P = 0.0001$. **d**, Suppressive effect of [D-Lys-3]-GHRP-6 (5 nmol) on feeding induced by ghrelin (20 pmol). Asterisk, $P = 0.012$.

reduction, implying that there is a competitive interaction between ghrelin and leptin in feeding regulation. We conclude that ghrelin is a physiological mediator of feeding, and probably has a function in growth regulation by stimulating feeding and release of growth hormone.

Ghrelin increased the food intake of rats in both satiated and feeding conditions. Intracerebroventricular (ICV) administration of ghrelin above a minimally active dose of 10 pmol to free-feeding rats during the early light phase (satiated) increased food intake in a dose-dependent manner (Fig. 1a). Ghrelin-treated rats showed no unusual behaviour relative to the controls. Administration of ghrelin also significantly increased dark phase (feeding) food intake (Fig. 1b). In rats that had fasted for 8 h, ghrelin also increased their 2-h food intake relative to the saline-injected group (Fig. 1c).

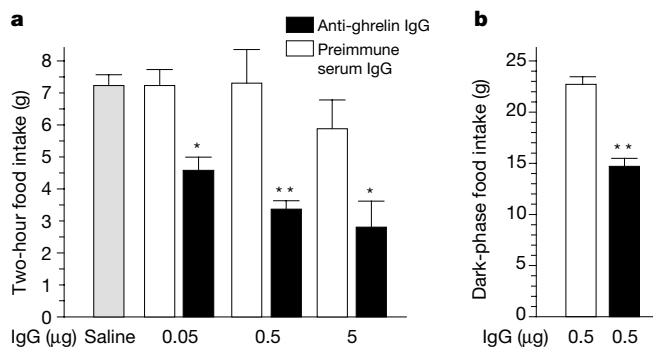


Figure 2 Anti-ghrelin IgG suppresses feeding. **a**, Food intake of rats that had fasted for 8 h, then received ICV administration of 0.05, 0.5 or 5 µg anti-ghrelin IgG or preimmune serum IgG at 08:45. **b**, Dark-phase food intake of free-feeding rats that received ICV administration of 0.5 µg anti-ghrelin IgG or preimmune serum IgG at 17:00. Asterisk, $P < 0.005$; double asterisk, $P < 0.0001$ versus preimmune IgG.

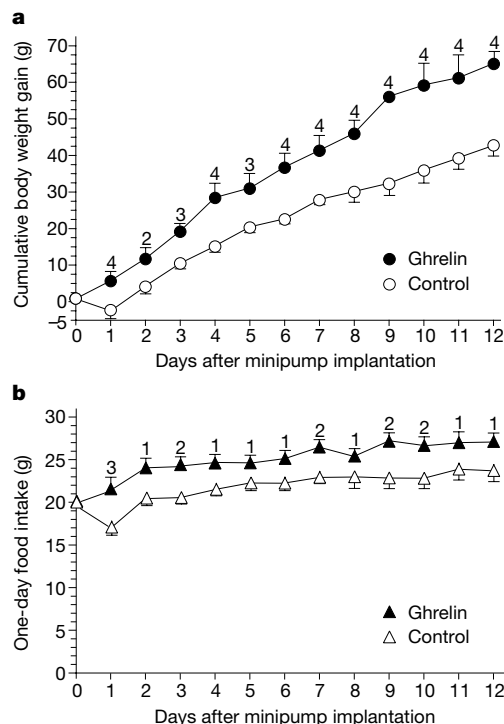


Figure 3 Effect of chronic ghrelin ICV administration on rats. Cumulative body weight gain (**a**) and one-day food intake (**b**) during an ICV infusion of 250 pmol d⁻¹ for 12 d. Alzet minipumps were implanted on day 0. 1, $P < 0.05$; 2, $P < 0.01$; 3, $P < 0.005$; 4, $P < 0.001$.

Centrally administered ghrelin induced feeding behaviour within 5 min of administration. GHRP-6 (modelled after enkephalins) is a synthetic hexapeptide that binds to growth hormone secretagogue receptor (GHS-R), releases growth hormone⁹ and stimulates feeding¹⁰. ICV-injected ghrelin was more effective at stimulating food intake than GHRP-6 (Fig. 1a). Ghrelin-induced feeding was suppressed by an antagonist for GHS-R¹¹, [D-Lys-3]-GHRP-6 (Fig. 1d).

To determine whether an endogenous tone of ghrelin signalling is present in the hypothalamus, we investigated the effect of an antibody against ghrelin on feeding behaviour. Compared with the preimmune serum immunoglobulin G (IgG), anti-ghrelin IgG suppressed starvation-induced feeding in a marked, dose-dependent manner (Fig. 2a). Anti-ghrelin IgG also suppressed dark phase food intake by 36% in free-feeding rats (Fig. 2b). These findings indicate that ghrelin is a powerful, endogenous orexigenic peptide.

A chronic ICV infusion of ghrelin (250 pmol d⁻¹) for 12 d using an osmotic minipump increased food intake and body weight gain over the infusion period (Fig. 3). It did not affect general activity (ghrelin: dark phase, 96 ± 6% of control activity; light phase, 95 ± 8%; $P \approx 0.5$), indicating that ghrelin does not mediate nonspecific arousal. The plasma concentrations of glucose, insulin, triglycerides and total cholesterol in the ghrelin-infused group did not differ from those in the control group (data not shown).

ICV-injected ghrelin also stimulated food intake in spontaneous dwarf rats (SDR)¹², a growth-hormone-deficient rat model that carries a disrupted growth-hormone gene¹³ (food intake: 200 pmol ghrelin, 1.18 ± 0.10 g; vehicle, 0.01 ± 0.01 g; $P < 0.0001$). Thus, the stimulatory effect of ghrelin on feeding does not depend on the stimulation of growth hormone.

To establish the neuronal populations activated by central ghrelin, we mapped *c-fos* expression after an ICV administration of ghrelin. Fos-immunoreactive neurons were observed primarily in regions implicated in the regulation of feeding behaviour (Fig. 4). This distribution is coincident with that of GHS-R⁴, which is also

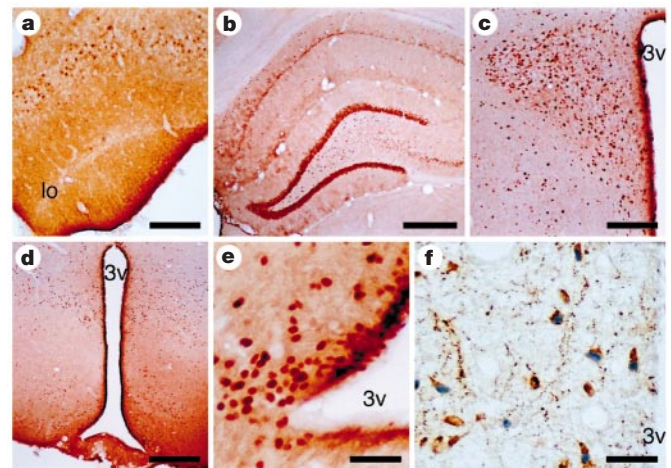


Figure 4 Localization of Fos expression in response to ICV administration of ghrelin. **a**, Piriform cortex. **b**, Dentate gyrus and hippocampus. **c**, Paraventricular nucleus. **d**, Arcuate, dorsomedial and ventromedial hypothalamic nuclei. **e**, High magnification of Fos immunoreactivity in the arcuate nucleus (magn. × 200). **f**, Co-staining of Fos (blue-black) and NPY neurons (brown) in the arcuate nucleus. All sections are from a rat given 0.5 nmol ghrelin. Fos is also found in the olfactory nerve layer; granular cell layer of the olfactory bulb; insular, prelimbic, infralimbic, orbital and cingulate cortices; accumbens, lateral septal, paraventricular thalamic, periventricular hypothalamic, anterior hypothalamic, supraoptic, suprachiasmatic, tuberomammillary, supramammillary and dorsal raphe nuclei (data not shown). lo, lateral olfactory tract; 3v, third ventricle. Scale bars: **a**, **c**, 200 µm; **b**, **d**, 500 µm; **e**, **f**, 50 µm. No Fos immunoreactivity is present in any of the regions observed in the control rats (data not shown).

named ghrelin receptor. Fos also was highly expressed in the dentate gyrus and hippocampus (Fig. 4b; CA1, CA2 and CA3) where ghrelin receptor messenger RNA is abundantly present. The possible involvement of ghrelin in learning and memory requires further investigation. No significant Fos expression was found in the neocortex nor in the cerebellum. Fos distributions were similar in the 0.01, 0.5 and 2 nmol ghrelin-injected rats (data not shown).

The arcuate nucleus is critical for feeding and body weight regulation because it has the leptin-responsive orexigenic neuropeptides, NPY^{14,15} and AGRP¹⁶, and the leptin-responsive anorexic neuropeptides, pro-opiomelanocortin¹⁷, and cocaine- and amphetamine-regulated transcript¹⁸. Of the three rats examined by double immunohistochemistry, ghrelin administration induced Fos expression in $39 \pm 6\%$ of NPY neurons in the medial part of the

arcuate nuclei (Fig. 4e, f), which is consistent with previous findings that NPY neurons have GHS receptors¹⁹ and express Fos in response to GHS administration^{20,21}.

We investigated the functional relationship between ghrelin and NPY by blocking either of the peptides in ghrelin- or NPY-induced feeding. Y1 and Y5 receptors are involved in feeding regulation by NPY^{22,23}. We first determined the doses of anti-NPY IgG and two antagonists for Y1 and Y5 receptors that are needed to block NPY-induced feeding, while inducing no other unusual behaviour (Fig. 5a). ICV administration of 1 μ g anti-NPY IgG 4 h before ghrelin administration cancelled ghrelin-induced feeding; co-administration of two antagonists for Y1 and Y5 receptors also cancelled ghrelin-induced feeding (Fig. 5b). In contrast, anti-ghrelin IgG did not affect NPY-induced feeding (Fig. 5a). Because

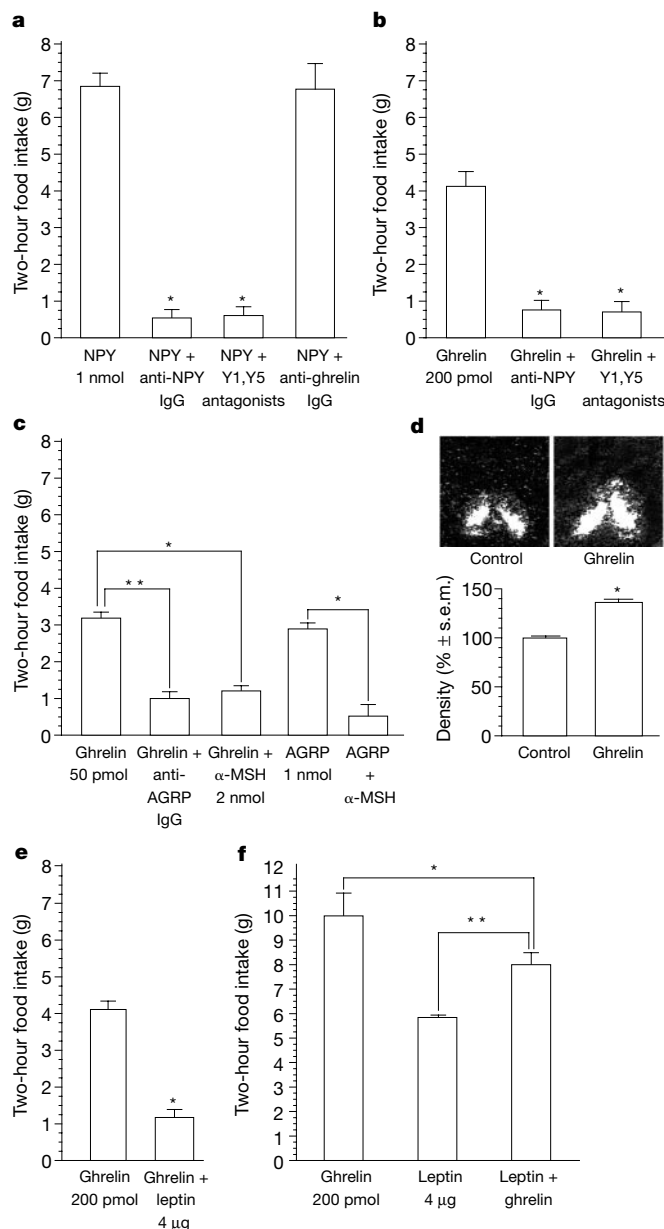


Figure 5 Interactions of ghrelin with NPY, AGRP and leptin. **a**, Effects of administration of anti-NPY IgG, anti-ghrelin IgG or co-administration of Y1 and Y5 antagonists on NPY-induced (1 nmol) feeding. Asterisk, $P < 0.0001$ versus NPY-injected group. **b**, Effects of anti-NPY IgG and co-administration of Y1 and Y5 antagonists on ghrelin-induced (200 pmol) feeding. Asterisk, $P < 0.0001$ versus ghrelin-injected group. **c**, Effects of anti-AGRP IgG or α -MSH on ghrelin-induced feeding. Asterisk, $P < 0.001$; double asterisk, $P < 0.0005$. **d**, *In situ* hybridization of hypothalamic NPY mRNA in rats

($n = 8$ per group) receiving ICV administration of ghrelin (1 nmol) or control vehicle. The quantitative image analysis of NPY mRNA expression in the arcuate nucleus is shown (bottom). Asterisk, $P < 0.01$. **e**, Suppressive effect of leptin on ghrelin-induced (200 pmol) feeding. Asterisk, $P < 0.0001$. **f**, Inhibitory effect of ghrelin on leptin-induced feeding reduction. Rats that had fasted for 8 h received an ICV administration of ghrelin (200 pmol), leptin (4 μ g) or leptin followed 1 h later by ghrelin. Asterisk, $P = 0.031$; double asterisk, $P = 0.016$.

AGRP co-localizes with NPY in arcuate nucleus neurons^{7,16}, we studied the relationship between ghrelin and AGRP in feeding regulation. Ghrelin-induced feeding was suppressed on both treatment with α -melanocyte-stimulating hormone (α -MSH), a melanocortin receptor agonist, and blocking of AGRP, a receptor antagonist²⁴, with anti-AGRP IgG (Fig. 5c). These results indicate that inhibition of endogenous NPY and AGRP may modulate ghrelin-induced feeding, which suggests that ghrelin interacts anatomically and/or functionally with the pathways of these two peptides.

Inhibition of NPY synthesis and release is a chief mechanism of food-intake reduction mediated by leptin^{25,26}. The hypothalamic NPY mRNA level in quantitative *in situ* hybridization²⁷ increased after ghrelin administration (Fig. 5d). Ghrelin-induced feeding in the light phase was suppressed by an ICV administration of leptin (Fig. 5e). Leptin reduced feeding in fasted rats, whereas ghrelin substantially blocked this reduction in rats that were pretreated with leptin (Fig. 5f). These results indicate that ghrelin may antagonize leptin action in the regulation of the NPY system.

Central ghrelin is a new physiological regulator of nutritional homeostasis. The classic effects of growth hormone in promoting growth of soft tissue, such as bone and cartilage, together with the orexigenic effect of ghrelin suggest that central and peripheral factors activated by ghrelin may underlie growth processes in an integrated manner. Further investigations of ghrelin's function will help our understanding of physiological feeding mechanisms and should facilitate the study of eating disorders.

Methods

Animals

We maintained male Wistar rats under controlled temperature and light conditions (light on 07:00–19:00). We performed cannulation and ICV administration as described²⁸. We repeated all of the experiments two or three times. All the compounds were dissolved in 0.9% saline, and 10 μ l solution in total was administered. We performed all procedures in accordance with the Japanese Physiological Society's guidelines for animal care.

Feeding experiments

First, various doses of rat ghrelin (Peptide Institute), GHRP-6 (Phoenix Pharmaceuticals) or ghrelin + [D-Lys-3]-GHRP-6 (Peninsula Laboratories) were administered by ICV injections to rats ($n = 16$ –20 per group) weighing 300–325 g at 08:45. Ghrelin (200 pmol) also was administered by ICV injection at 08:45 to rats ($n = 12$) that had fasted for 8 h. We re-weighed chow 2 h after peptide administration, and calculated food intake. Second, ghrelin (200 pmol) or saline was administered by ICV injection at 18:45 to free-feeding rats ($n = 12$ per group), after which dark phase (19:00–07:00) food intake was measured. Third, ghrelin (250 pmol per 14 μ l saline per day, for 12 d) or vehicle was infused continuously through osmotic minipumps to 7-week-old Wistar rats ($n = 10$ per group). Cannulae implanted into the lateral ventricles were connected to minipumps (Alzet, type 2002) inserted under the skin of the neck. We measured body weight and food consumption daily at 07:00. On day 12, the rats were killed, and the trunical blood was sampled. Fourth, ghrelin (200 pmol) was administered by ICV injection to 16-week-old spontaneous dwarf rats (SDR) ($n = 6$) weighing 95–100 g (Japan SLC), after which 2-h food intake was measured.

To investigate the functional relationship between ghrelin and NPY or AGRP, ghrelin was co-administered with an antagonist or antibody for either of the peptides. Rats ($n = 12$ –16 per group) were administered an ICV injection in the morning with the following reagents: ghrelin (200 pmol); ghrelin + anti-NPY IgG (1 μ g; Peptide Institute); ghrelin + 1229U91 (30 μ g, a Y1 antagonist²⁹) + L-152,804 (30 μ g, a selective Y5 antagonist³⁰); NPY (1 nmol); NPY + 1229U91 + L-152,804; NPY + anti-ghrelin IgG (0.5 μ g); NPY + anti-NPY IgG; ghrelin + anti-AGRP IgG (1 μ g, Phoenix Pharmaceuticals); ghrelin + α -MSH; AGRP (1 nmol) or AGRP + α -MSH (2 nmol). Doses of each reagent used were the same among the sets except for ghrelin (20, 50 and 200 pmol) + α -MSH (1, 2 and 4 nmol). We injected IgG 4 h before peptide administration in all cases. We monitored food intake for 2 h.

We conducted two experiments to study the interaction between ghrelin and leptin in feeding regulation. First, free-feeding rats ($n = 12$ per group) were administered an ICV injection in the morning with ghrelin (200 pmol) or ghrelin + mouse leptin (4 μ g, a gift from the National Hormone and Pituitary Program). Second, rats ($n = 12$ per group) that had fasted for 8 h were administered an ICV injection in the morning with ghrelin (200 pmol), leptin (4 μ g) or ghrelin + leptin. We injected leptin at 07:45 and other peptides at 08:45 in both of the experiments. We measured 2-h food intake. We analysed groups of data (mean $\pm$ s.e.m.) using ANOVA (analysis of variance) and *post hoc* Fisher's test.

Immunoneutralization

We subjected anti-ghrelin antiserum¹ to Affi-gel protein A affinity and then CNBr-Sephacryl-coupled ghrelin affinity chromatography. We determined the amount of

purified IgG by using a DC protein assay kit (Bio-Rad). First, a 10 μ l saline solution of purified anti-ghrelin IgG or preimmune serum IgG from the same rabbit was administered by ICV injection at 08:45 the following morning to rats ($n = 10$ per group) that had fasted for 8 h. Second, IgG was given through ICV to free-feeding rats ($n = 10$ per group) at 17:00. We measured food intake after IgG administration.

Locomotor activity

Movement of rats ($n = 10$ per group) that had been given a continuous ICV infusion of ghrelin (250 pmol per 14 μ l saline, for 5 d) or the vehicle through osmotic minipumps was measured on days 1–5 as described³¹. We made locomotor-activity counts every 15 min and summed them for the dark and light phases.

c-fos expression

We studied four rat groups ($n = 3$ per group; 10 pmol ghrelin, 500 pmol ghrelin, 2 nmol ghrelin and 0.9% saline). We administered an ICV injection of ghrelin or saline 90 min before perfusion. Frozen serial brain sections (40 μ m thick) were incubated for 2 d with goat anti-c-Fos antiserum (Santa Cruz Biotechnology; final dilution 1:1,500)³². We stained the sections by the avidin–biotin complex method³². We subjected some sections of the arcuate nucleus to Fos staining, and then to double staining with rabbit anti NPY-antiserum (DiaSorin; final dilution 1:4,000).

In situ hybridization

We performed *in situ* hybridization of NPY mRNA with a 45-nucleotide antisense probe as described²⁷. We analysed the images in an MCID imaging analyser³⁷.

Received 9 August; accepted 2 November 2000.

- Kojima, M. *et al.* Ghrelin is a growth-hormone-releasing acylated peptide from stomach. *Nature* **402**, 656–660 (1999).
- Howard, A. D. *et al.* A receptor in pituitary and hypothalamus that functions in growth hormone release. *Science* **273**, 974–977 (1996).
- McKee, K. K. *et al.* Molecular analysis of rat pituitary and hypothalamic growth hormone secretagogue receptors. *Mol. Endocrinol.* **11**, 415–423 (1997).
- Guan, X. M. *et al.* Distribution of mRNA encoding the growth hormone secretagogue receptor in brain and peripheral tissues. *Mol. Brain Res.* **48**, 23–29 (1997).
- Sagar, S. M., Sharp, F. R. & Curran, T. Expression of c-fos protein in brain: metabolic mapping at the cellular level. *Science* **240**, 1328–1331 (1988).
- Stanley, B. G., Kyrikouli, S. E., Lampert, S. & Leibowitz, S. F. Neuropeptide Y chronically injected into the hypothalamus: a powerful neurochemical inducer of hyperphagia and obesity. *Peptides* **7**, 1189–1192 (1986).
- Hahn, T., Breininger, J., Baskin, D. & Schwartz, M. Coexpression of AgRP and NPY in fasting-activated hypothalamic neurons. *Nature Neurosci.* **1**, 271–272 (1998).
- Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–432 (1994).
- Bowers, C. Y., Momany, F. A., Reynolds, G. A. & Hong, A. On the *in vitro* and *in vivo* activity of a new synthetic hexapeptide that acts on the pituitary to specifically release growth hormone. *Endocrinology* **114**, 1537–1545 (1984).
- Locke, W., Kirgis, H. D., Bowers, C. Y. & Abdoh, A. A. Intracerebroventricular growth-hormone-releasing peptide-6 stimulates eating without affecting plasma growth hormone responses in rats. *Life Sci.* **56**, 1347–1352 (1995).
- Cheng, K. *et al.* The synergistic effects of His-D-Trp-Ala-Trp-D-Phe-Lys-NH₂ on growth hormone (GH)-releasing factor-stimulated GH release and intracellular adenosine 3',5'-monophosphate accumulation in rat primary pituitary cell culture. *Endocrinology* **124**, 2791–2798 (1989).
- Okuma, S. & Kawashima, S. Spontaneous dwarf rat. *Exp. Anim.* **29**, 301–304 (1980).
- Takeuchi, T. *et al.* Molecular mechanism of growth hormone (GH) deficiency in the spontaneous dwarf rat: detection of abnormal splicing of GH messenger ribonucleic acid by the polymerase chain reaction. *Endocrinology* **126**, 31–38 (1990).
- Schwartz, M. W. *et al.* Identification of targets of leptin action in rat hypothalamus. *J. Clin. Invest.* **98**, 1101–1106 (1996).
- Mercer, J. G. *et al.* Coexpression of leptin receptor and prepro-neuropeptide Y mRNA in arcuate nucleus of mouse hypothalamus. *J. Neuroendocrinol.* **8**, 733–735 (1996).
- Broberger, C. *et al.* The neuropeptide Y/agouti gene-related protein (AGRP) brain circuitry in normal, anorectic, and monosodium glutamate-treated mice. *Proc. Natl Acad. Sci. USA* **95**, 15043–15048 (1998).
- Schwartz, M. W. *et al.* Leptin increases hypothalamic pro-opiomelanocortin mRNA expression in the rostral arcuate nucleus. *Diabetes* **46**, 2119–2123 (1997).
- Kristensen, P. *et al.* Hypothalamic CART is a new anorectic peptide regulated by leptin. *Nature* **393**, 72–76 (1998).
- Willemsen, M. G., Kristensen, P. & Romer, J. Co-localization of growth hormone secretagogue receptor and NPY mRNA in the arcuate nucleus of the rat. *Neuroendocrinology* **70**, 306–316 (1999).
- Dickson, S. L. & Luckman, S. M. Induction of c-fos messenger ribonucleic acid in neuropeptide Y and growth hormone (GH)-releasing factor neurons in the rat arcuate nucleus following systemic injection of the GH secretagogue, GH-releasing peptide-6. *Endocrinology* **138**, 771–777 (1997).
- Honda, K. *et al.* An electrophysiological and morphological investigation of the projections of growth hormone-releasing peptide-6-responsive neurons in the rat arcuate nucleus to the median eminence and to the paraventricular nucleus. *Neuroscience* **90**, 875–883 (1999).
- Gerald, C. *et al.* A receptor subtype involved in neuropeptide-Y-induced food intake. *Nature* **382**, 168–171 (1996).
- Wieland, H. A. *et al.* Subtype selectivity of the novel nonpeptide neuropeptide Y Y1 receptor antagonist BIBO 3304 and its effect on feeding in rodents. *Br. J. Pharmacol.* **125**, 549–555 (1998).
- Cone, R. D. *et al.* The melanocortin receptors: agonists, antagonists, and the hormonal control of pigmentation. *Rec. Prog. Horm. Res.* **51**, 287–317 (1996).
- Stephens, T. W. *et al.* The role of neuropeptide Y in the antiobesity action of the obese gene product. *Nature* **377**, 530–532 (1995).

26. Schwartz, M. W. *et al.* Specificity of leptin action on elevated blood glucose levels and hypothalamic neuropeptide Y gene expression in *ob/ob* mice. *Diabetes* **45**, 531–535 (1996).
27. Yamamoto, Y. *et al.* Down regulation of the prepro-orexin gene expression in genetically obese mice. *Mol. Brain Res.* **65**, 14–22 (1999).
28. Ida, T., Nakahara, K., Murakami, N. & Nakazato, M. Effects of lateral cerebroventricular injection of the appetite-stimulating neuropeptides, orexin and neuropeptide Y, on the various behavioral activities of rats. *Brain Res.* **821**, 526–529 (1999).
29. Daniels, A. J. *et al.* High-affinity neuropeptide Y receptor antagonists. *Proc. Natl Acad. Sci. USA* **92**, 9067–9071 (1995).
30. Kanatani, A. *et al.* L-152,804: orally active and selective neuropeptide Y Y5 receptor antagonist. *Biochem. Biophys. Res. Commun.* **272**, 169–173 (2000).
31. Murakami, N., Marumoto, N., Nakahara, K. & Murakami, T. Daily injections of melatonin entrain the circadian activity rhythms of nocturnal rats but not diurnal chipmunks. *Brain Res.* **775**, 240–243 (1997).
32. Date, Y. *et al.* Orexin, orexigenic hypothalamic peptides, interact with autonomic, neuroendocrine and neuroregulatory systems. *Proc. Natl Acad. Sci. USA* **96**, 748–753 (1999).

Acknowledgements

We thank Y. Ueta for *in situ* hybridization; T. Kuroiwa, Y. Kawabata and R. Matsuura for assistance; and M. Ihara and A. Kanatani for providing L-152,804. This work was supported in part by grants-in-aid from the Ministry of Education, Science, Sports and Culture, and the Ministry of Health and Welfare, Japan, to M.N.

Correspondence and requests for materials should be addressed to M.N.
(e-mail: nakazato@post.miyazaki-med.ac.jp).

Identification of the haemoglobin scavenger receptor

Mette Kristiansen^{*†}, Jonas H. Graversen^{*†}, Christian Jacobsen^{*},
Ole Sonne[‡], Hans-Jürgen Hoffman[§], S.K. Alex Law^{||}
& Søren K. Moestrup^{*}

^{*} Department of Medical Biochemistry, and [‡] Department of Physiology, University of Aarhus, Ole Worms Alle, 8000 Aarhus C, Denmark

[§] Department of Respiratory Diseases, Aarhus University Hospital, 8000 Aarhus C, Denmark

^{||} MRC Immunochemistry Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

[†] These authors contributed equally to the work

Intravascular haemolysis is a physiological phenomenon as well as a severe pathological complication when accelerated in various autoimmune, infectious (such as malaria) and inherited (such as sickle cell disease) disorders¹. Haemoglobin released into plasma is captured by the acute phase protein haptoglobin, which is depleted from plasma during elevated haemolysis¹. Here we report the identification of the acute phase-regulated and signal-inducing macrophage protein, CD163, as a receptor that scavenges haemoglobin by mediating endocytosis of haptoglobin-haemoglobin complexes. CD163 binds only haptoglobin and haemoglobin in complex, which indicates the exposure of a receptor-binding neopeptide. The receptor-ligand interaction is Ca²⁺-dependent and of high affinity. Complexes of haemoglobin and multimeric haptoglobin (the 2-2 phenotype) exhibit higher functional affinity for CD163 than do complexes of haemoglobin and dimeric haptoglobin (the 1-1 phenotype). Specific CD163-mediated endocytosis of haptoglobin-haemoglobin complexes is measurable in cells transfected with CD163 complementary DNA and in CD163-expressing myelo-monocytic lymphoma cells.

Metabolism of haemoglobin (Hb), the most abundant protein in erythrocytes and blood, is a main function of tissue macrophages, which can engulf senescent erythrocytes (extravascular haemolysis) or take up haemoglobin released from ruptured erythrocytes (intravascular haemolysis) and immature erythrocytes in the bone marrow. Efficient removal of free Hb is essential for health because of the oxidative and toxic properties of the iron-containing haem in Hb. In the macrophage, haem is converted to bilirubin and iron¹.

Whereas various receptors^{2–8} for direct or indirect binding of surface-exposed phosphatidyl-serine are suggested to be involved in the recognition and uptake of senescent erythrocytes, the cellular structure involved in the clearance of plasma Hb by macrophages has remained unknown. The plasma protein haptoglobin (Hp) is thought to be involved in promoting the clearance of plasma Hb⁹, because it strongly binds free Hb and is depleted during elevated haemolysis¹. To identify the molecular recognition events determining the clearance of Hb released into plasma during intravascular haemolysis, we constructed an Hp-Hb affinity matrix for the purification of a putative receptor for this complex. Subsequent affinity chromatography of solubilized membranes from three macrophage-containing human tissues (placenta, liver and spleen) yielded a protein with relative molecular mass of 130,000 (*M_r* 130K) (Fig. 1a). Matrix-assisted laser-desorption/ionization (MALDI) mass spectrometry of a tryptic digest of the 130K protein (Fig. 1b) identified it as the scavenger receptor cysteine-rich domain protein, M130/CD163 (refs 10–12), which is an acute phase-regulated transmembrane protein that is expressed exclusively in monocytes (low expression) and tissue macrophages (high expression)¹³. Consistent with the difference in macrophage content of the source tissues for the affinity chromatography, the highest yield was obtained from the spleen (~0.1–0.2 mg CD163 per g membrane). The yields from liver and placenta were about 4 times and 20 times lower, respectively. Isolated CD163 from any of the tissues was of very high purity, and no other proteins, including liver- or placenta-specific proteins, were detected at significant levels. We identified a protein that had an electrophoretic mobility identical to that of CD163 by ¹²⁵I-labelled Hp-Hb blotting of solubilized spleen membranes (Fig. 1c, lane 4). We confirmed the identity of the Hp-Hb-binding protein as CD163 by immunoblotting (Fig. 1c, lanes 5–8) with two different monoclonal antibodies against CD163 (refs 13 and 14).

Haptoglobin is synthesized as a single chain, which is cleaved to an amino-terminal α -chain and a carboxy-terminal β -chain. The basic structure of Hp, as found in most mammals, is a homodimer (Fig. 2a) designated Hp(1-1) in which the two Hp molecules are linked by a single disulphide bond through their respective ~9K α -chains¹⁵. In humans, a variant with a longer α -chain is also present in all populations. This variant arose apparently by an early intragenic duplication, presumably originating from an unequal crossover of two basic alleles, resulting in an Hp with an α -chain of ~14K. The short and long α -chains are designated as α^1 and α^2 , respectively. As the cysteine forming the intermolecular disulphide bond between the α -chains is also duplicated, humans homozygous for the long variant allele show a multimeric Hp phenotype (Fig. 2a) designated Hp(2-2). Hp(2-1) refers to the phenotype (both Hp dimers and multimers) seen in humans heterozygous for the two variant alleles.

Analysis of Hp-Hb complexes binding to immobilized CD163 showed a high-affinity binding of both dimeric and multimeric Hp-Hb complexes (Fig. 2b, c). Figure 2b shows a surface-plasmon resonance analysis of CD163 binding of the dimeric Hp(1-1)-Hb complex and the multimeric Hp(2-2)-Hb complex. No binding of non-complexed Hb (Fig. 2b, left panel), Hp(1-1) (Fig. 2b, middle panel) nor Hp(2-2) (Fig. 2b, right panel) was detected, thus indicating that a neopeptide for receptor binding is exposed in the Hp-Hb complex. Accordingly, maximal receptor binding was measured when the Hb-binding capacity of Hp reached saturation at equimolar concentrations of Hb and Hp (Fig. 2b, middle and right panels). The Hp(2-2)-Hb complex yielded a higher response and the dissociation was slower as compared with the Hp(1-1)-Hb complex. The results shown in Fig. 2b were obtained using the A₀ ($\alpha 2\beta 2$) form of Hb. We obtained similar results using the A₂ ($\alpha 2\delta 2$) form, or the S form (Hb with the mutation for sickle-cell disease)¹⁶ (data not shown).

We used a solid-phase assay with immobilized CD163 in micro-

titre wells for binding inhibition experiments (Fig. 2c). This analysis showed that the removal of Ca^{2+} with EDTA or the addition of polyclonal anti-CD163 immunoglobulin- γ (IgG) abolished completely the binding of Hp-Hb to CD163. Measuring the true affinity of the one-site interaction of Hp-Hb binding to CD163 was hampered by the suggested divalency (Hp(1-1)) and multivalency (Hp(2-2)) of the ligand in terms of receptor-recognition sites. However, competition for CD163-binding of ^{125}I -labelled Hp-Hb by unlabelled Hp(1-1)-Hb and Hp(2-2)-Hb complexes showed, as anticipated from the surface-plasmon resonance experiments, a roughly 10-fold increase in functional affinity (avidity) of the multimeric Hp(2-2)-Hb complexes (Fig. 2c). The concentration of unlabelled Hp(1-1)-Hb complex causing 50% inhibition of the binding of ^{125}I -labelled Hp(1-1)-Hb was about $0.3 \mu\text{g ml}^{-1}$, giving an 'apparent dissociation constant' (K_d) of about 2 nM of the dimeric Hp(1-1)-Hb complex. The 50% inhibition point for Hp(2-2)-Hb was at about $0.1 \mu\text{g ml}^{-1}$ giving an apparent K_d of about 0.2 nM (on assumption of the distribution of the various 2-2 forms previously calculated¹⁵). The higher functional affinity of the 2-2 type complex is probably accounted for by its higher valency. A similar 'bonus effect of multivalency' is well known in other biological systems, such as the binding of the pentameric IgM molecule to several identical surface antigens.

CD163-mediated endocytosis of ^{125}I -labelled Hp-Hb complexes was studied in Chinese hamster ovary (CHO) cells transfected with CD163 cDNA (the most abundant CD163 form). Figure 3a (middle panel) shows the time course of cell-associated radioactivity and trichloroacetic acid (TCA)-soluble radioactivity (representing degraded ligand) in the medium. The cell-associated radioactivity

reached a plateau after 1 h of incubation, at about the same time as the TCA-soluble radioactivity increased significantly in the medium. A similar experiment consistent with an endocytic uptake of Hp-Hb, conducted in the presence of the lysosomal inhibitors chloroquine and leupeptin, showed a continual increase in cell-bound radioactivity for 3 h with essentially no TCA-soluble radioactivity detected (Fig. 3a, right panel).

The endocytosis of Hp-Hb complexes was mediated by CD163, as no uptake and consequently no TCA-soluble radioactivity was detected in incubations with CHO cells that did not express CD163 (Fig. 3a, left panel). Furthermore, uptake and degradation of ^{125}I -labelled Hp(2-2)-Hb were inhibited by purified IgG from anti-CD163 serum and by unlabelled Hp(2-2)-Hb complexes (Fig. 3b, left panel). We obtained similar results with ^{125}I -labelled Hp(1-1)-Hb complexes, although a lower rate of uptake was observed in comparison with the ^{125}I -labelled Hp(2-2)-Hb complexes (data not shown). We also measured uptake of ^{125}I -labelled Hp(2-2)-Hb (Fig. 3b, right panel) in the myelo-monocytic¹⁷ SU-DHL-1 cells that express CD163¹². This cell line expresses, in addition to the most abundant CD163 variant, two less abundant variants^{11,12} with different cytoplasmic tails. It remains to be clarified whether these three variants show functional differences. However, their identical extracellular domains suggest that they all bind Hp-Hb.

Previous work on the biological function of CD163 is limited to a study on the effect of antibody-mediated crosslinking of CD163 on cultured monocytes¹⁴. The CD163 surface ligation induces a tyrosine-kinase-dependent signal resulting in intracellular calcium mobilization, inositol triphosphate production and increased secretion of anti-inflammatory cytokines. Considering our results, we

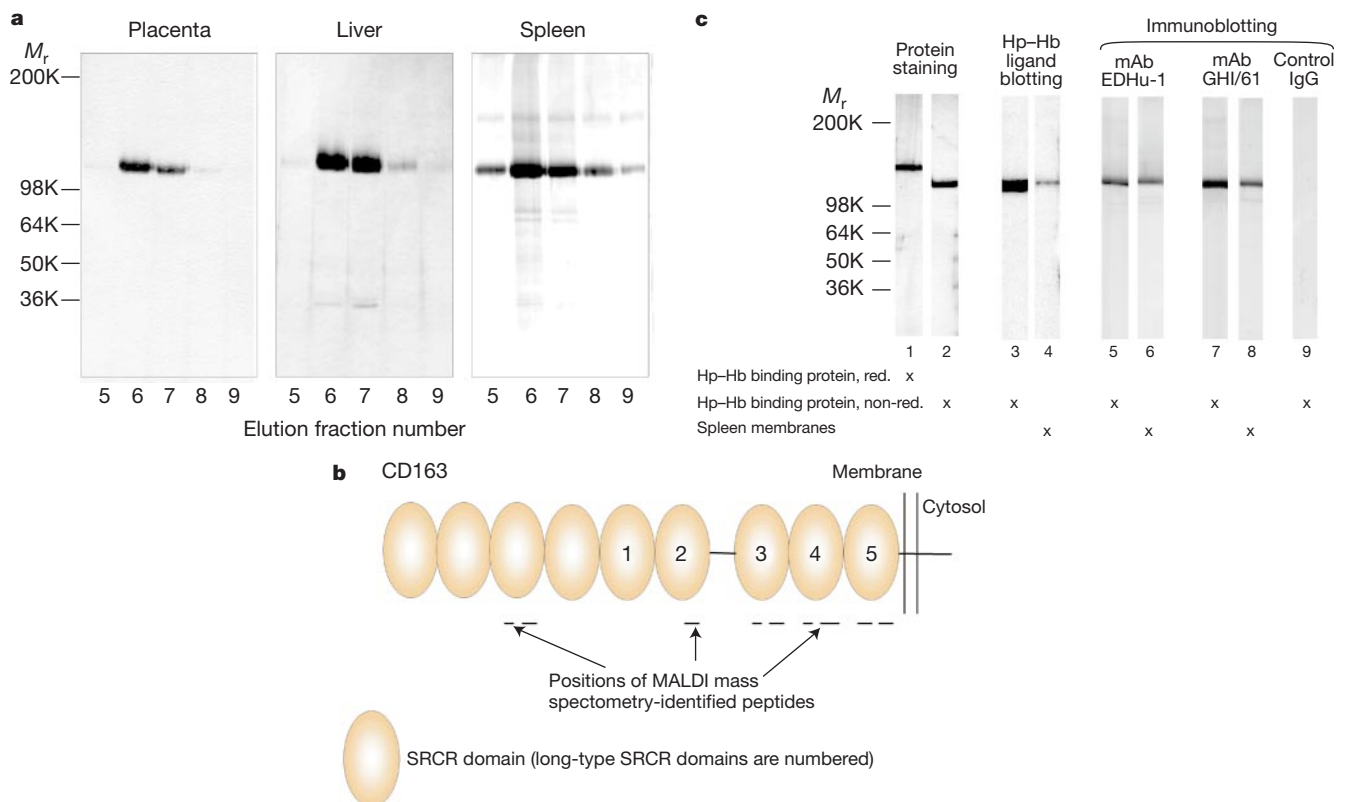


Figure 1 Identification of CD163 as a binding protein of Hp-Hb. **a**, Ligand-affinity chromatography of solubilized membranes from human placenta, liver and spleen. A Hp-Hb-binding protein was eluted with an EDTA-phosphate buffer (pH 6). SDS-PAGE showed a main band of ~130K. **b**, Identification of the protein by MALDI mass spectrometry. The mass of ten tryptic peptides corresponds to the mass of amino acids 301–313, 348–369, 633–642, 790–804, 834–847, 848–867, 893–907,

908–923, 976–992 and 1,008–1,016 of human CD163. **c**, SDS-PAGE of the ~130K protein (lanes 1, 2), ligand blotting with ^{125}I -labelled Hp-Hb complex (lanes 3, 4) and immunoblotting with the monoclonal antibodies EDHu-1 and GHI/61 (lanes 5–8) or control IgG (lane 9). The table below the gel indicates the nature of the protein samples (purified Hp-Hb binding 130 K protein or spleen membranes). The samples were electrophoresed at either reducing (red) or non-reducing (non-red) conditions.

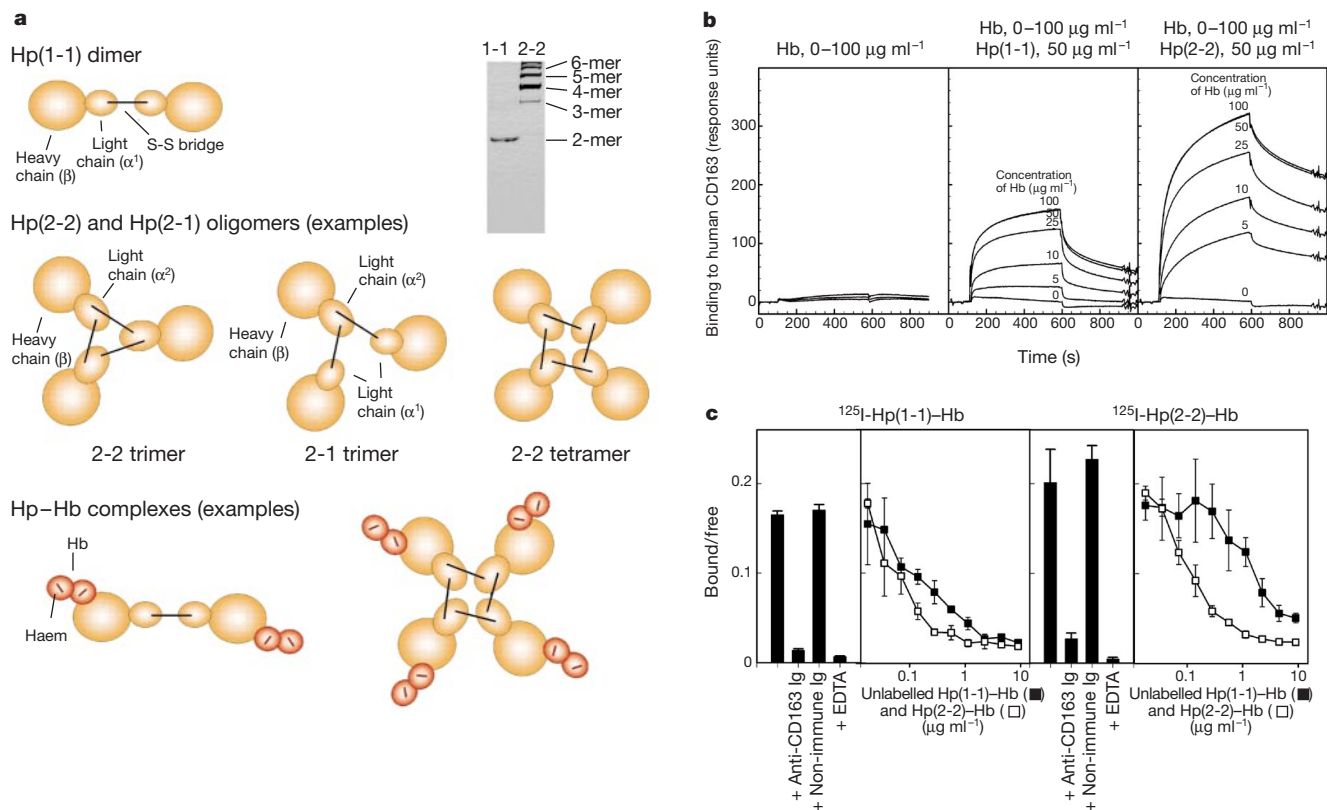


Figure 2 Binding of Hp-Hb to CD163. **a**, Subunit organization of Hp and Hp-Hb complexes. The inset shows non-reducing SDS-PAGE of the Hp(1-1) dimer and Hp(2-2) multimers. **b**, Surface-plasmon resonance analysis of binding of Hp-Hb to CD163. The measurements were carried out at Hb concentrations ranging from 0 to 100 $\mu\text{g ml}^{-1}$ in

the absence or presence of 50 $\mu\text{g ml}^{-1}$ of Hp(1-1) or Hp(2-2). No binding was observed with either Hb or Hp alone, and saturation of the binding was seen when Hb reached the same concentration as Hp. **c**, Inhibition of ^{125}I -labelled Hp(1-1)-Hb and Hp(2-2)-Hb binding to CD163 (immobilized in microtitre wells) by antibodies, EDTA (5 mM) or Hp-Hb.

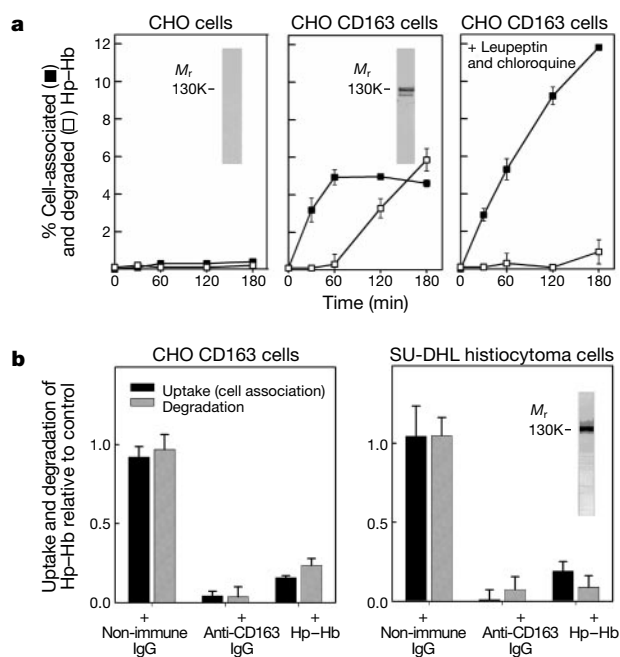


Figure 3 CD163-mediated endocytosis of ^{125}I -labelled Hp-Hb. **a**, Cell association and degradation of ^{125}I -labelled Hp(2-2)-Hb in mock-transfected CHO cells (left) and CHO cells transfected with CD163 cDNA (middle and right). Addition of the lysosomal inhibitors chloroquine and leupeptin (both 100 μM) inhibited degradation leading to cellular accumulation of radioactivity (right). **b**, Inhibition of ^{125}I -labelled Hp-Hb uptake in CHO

cells transfected with CD163 cDNA (left) and in histiocytic lymphoma-derived SU-DHL-1 cells that express CD163 (right). Both cell types show a saturable uptake that is inhibited by anti-CD163 polyclonal IgG. Insets in **a** and **b** show anti-CD163 immunoblotting of the cells.

speculate that the antibodies may mimic the crosslinking of CD163 by multimeric Hp–Hb. If so, multimeric Hp–Hb complexes may cause a stronger anti-inflammatory reaction as compared with dimeric Hp–Hb.

The expression of both CD163 and Hp is highly upregulated by the main acute phase reactants, interleukin (IL-6) and glucocorticoids^{10,18,19}. Haem oxygenase, an essential enzyme in the early metabolism of haem, is also upregulated by IL-6 (ref. 20). Thus, Hp, CD163 and haem oxygenase may be regarded as a 'capture molecule–receptor–enzyme' system that is coherently upregulated during inflammatory conditions to enhance the capacity for Hb clearance. In addition, the activation of this system may trigger the secretion of anti-inflammatory cytokines.

In its function as an Hb scavenger receptor, CD163 may account for a substantial transfer of iron into the macrophages, which has an integral function in iron homeostasis. A high turnover of iron in macrophages²¹ is consistent with the high expression of the iron exporter ferroportin-1 (ref. 22) in both Kupffer cells in the liver and macrophages in the red splenic pulp. Whether CD163 function may be perturbed in human disorders of abnormal iron homeostasis is unknown.

We have thus shown that binding of Hb to Hp leads to exposure of a neoepitope that promotes receptor recognition; that the receptor for uptake of Hp–Hb complexes is identical to the acute phase-regulated CD163 expressed exclusively in monocytes/macrophages; and that the multimerization of Hp leads to increased functional receptor affinity of the Hp–Hb complex. These data reveal the mechanism for Hb clearance, and provide new perspectives for understanding the potential immunological properties of Hp during inflammatory conditions. Furthermore, the recognition and endocytosis of Hp–Hb by CD163 is the first physiological function attributed to a protein in group B of the scavenger receptor cysteine-rich domain receptors²³. We propose the functional name 'haemoglobin scavenger receptor' (HbSR) for CD163. Future studies may disclose whether CD163/HbSR has additional ligands, such as other molecules produced by cell damage or inflammation. □

Methods

Purification and identification of the Hp–Hb receptor

We obtained human Hp (1-1, 2-2 and mixed phenotypes) and human Hb (A₀, A₂ and S forms) from Sigma. We prepared a 5-ml Hp–Hb Sepharose CL-4B (Pharmacia–Amersham) column by coupling complexes of Hp (5 mg, mixed phenotypes) and Hb (4 mg, type A₀). We loaded the column with 100 ml ~1% Triton X-100 solubilized membranes (from human spleen, placenta and liver) as described²⁴. The purified 130K protein binding Hp–Hb was eluted in 10 mM NaH₂PO₄, pH 6, 150 mM NaCl, 5 mM EDTA and 0.5% CHAPS (Aldrich). Protein separated by SDS–polyacrylamide gel electrophoresis was processed for tryptic digestion and MALDI mass spectrometry by Protana (Odense, Denmark). The difference in calculated and measured masses for all peptides was less than 0.042 daltons. We used the murine monoclonal CD163 antibodies EDHu-1 (Serotec) and GHI/61 (Research Diagnostics) for western blotting. We raised a polyclonal CD163 antibody by immunization of a rabbit with ligand-affinity purified receptor.

Ligand-receptor binding analysis

We performed surface-plasmon resonance analysis as described²⁵. Purified CD163 was immobilized at the BIAcore sensor CM5 chip (BIAcore AB) at a concentration of up to 50 µg ml⁻¹ in 10 mM sodium acetate, pH 4.0, and the remaining binding sites were blocked with 1 M ethanolamine, pH 8.5. The surface-plasmon resonance signal generated from immobilized CD163 corresponded to 55–66 fmol receptor per mm². The sample and flow buffer used was 10 mM HEPES, 150 mM NaCl and 0.5 mM CaCl₂ pH 7.4. We regenerated the sensor chips with 1.6 M glycine–HCl, pH 3. The binding assay for measuring binding of ¹²⁵I-labelled Hp–Hb to human CD163 immobilized in microtitre plate wells (Nunc) was performed as described²⁶. We coated the microtitre plates at 4 °C for 20 h with purified CD163 in 50 mM NaHCO₃ containing 250 ng CD163 per well (for binding ¹²⁵I-labelled Hp(1-1)–Hb) or containing 125 ng CD163 per well (for binding ¹²⁵I-labelled Hp(2-2)–Hb). We performed iodination of Hp–Hb with the chloramine-T method. We carried out ligand blotting as described²⁷, using 10⁶ c.p.m. radioligand per ml.

Endocytosis analysis

The cDNA encoding the most abundant variant of CD163 (ref. 11) was ligated into the *Kpn*I and *Not*I sites of the mammalian expression vector pcDNA3.1/Zeo(+) (Invitrogen). We established stable, transfected CHO clones expressing CD163 by limited dilution and

selection with 500 µg ml⁻¹ Zeocin (Invitrogen). We analysed expression products by immunoblotting of growth medium and cell lysate, using the rabbit polyclonal antibody against the ligand-affinity-purified human CD163.

Endocytosis of ¹²⁵I-labelled Hp–Hb in CD163-transfected and mock-transfected CHO cells growing as confluent adherent monolayers in 24-well plates was analysed as described²⁸. We analysed endocytosis in the non-adherent SU-DHL-1 histiocytic lymphoma cells (2 × 10⁶ cell ml⁻¹) as described²⁹.

Received 7 August; accepted 10 October 2000.

- Deiss, A. in *Wintrobe's Clinical Hematology* (eds Lee, G. R. et al.) 267–299 (Williams & Wilkins, Baltimore, 1999).
- Zwaal, R. F. & Schroit, A. J. Pathophysiologic implications of membrane phospholipid asymmetry in blood cells. *Blood* **89**, 1121–1132 (1997).
- Rigotti, A., Acton, S. L. & Krieger, M. The class B scavenger receptors SR-BI and CD36 are receptors for anionic phospholipids. *J. Biol. Chem.* **270**, 16221–16224 (1995).
- Krieger, M. et al. Molecular flypaper, host defense, and atherosclerosis. Structure, binding properties, and functions of macrophage scavenger receptors. *J. Biol. Chem.* **268**, 4569–4572 (1993).
- Devitt, A. et al. Human CD14 mediates recognition and phagocytosis of apoptotic cells. *Nature* **392**, 505–509 (1998).
- Oka, K. et al. Lectin-like oxidized low-density lipoprotein receptor 1 mediates phagocytosis of aged/apoptotic cells in endothelial cells. *Proc. Natl Acad. Sci. USA* **95**, 9535–9540 (1998).
- Sambrano, G. R., Parthasarathy, S. & Steinberg, D. Recognition of oxidatively damaged erythrocytes by a macrophage receptor with specificity for oxidized low density lipoprotein. *Proc. Natl Acad. Sci. USA* **91**, 3265–3269 (1994).
- Fadok, V. A. et al. A receptor for phosphatidylserine-specific clearance of apoptotic cells. *Nature* **405**, 85–90 (2000).
- Wada, T., Oara, H., Watanabe, K., Kinoshita, H. & Yachi, A. Autoradiographic study on the site of uptake of the haptoglobin-haemoglobin complex. *J. Reticuloendothel. Soc.* **8**, 185–193 (1970).
- Buechler, C. et al. Regulation of scavenger receptor CD163 expression in human monocytes and macrophages by pro- and antiinflammatory stimuli. *J. Leukoc. Biol.* **67**, 97–103 (2000).
- Law, S. K. et al. A new macrophage differentiation antigen which is a member of the scavenger receptor superfamily. *Eur. J. Immunol.* **23**, 2320–2325 (1993).
- Pulford, K., Micklem, K., Law, S. K. & Mason, D. Y. in *Leukocyte Typing VI*. (eds Kishimoto, T. et al.) 1089–1091 (Garland, New York, 1997).
- Pulford, K. et al. A monocyte/macrophage antigen recognized by the four antibodies GHI/61, BEMAC3, Ki-M8 and SM4. *Immunology* **75**, 588–595 (1992).
- Van den Heuvel, M. M. et al. Regulation of CD163 on human macrophages: cross-linking of CD163 induces signaling and activation. *J. Leukoc. Biol.* **66**, 858–866 (1999).
- Wejman, J. C., Hovsepian, D., Wall, J. S., Hainfeld, J. F. & Greer, J. Structure and assembly of haptoglobin polymers by electron microscopy. *J. Mol. Biol.* **174**, 343–368 (1984).
- Bunn, H. F. Pathogenesis and treatment of sickle cell disease. *N. Engl. J. Med.* **337**, 762–769 (1997).
- Epstein, A. L. et al. Biology of the human malignant lymphomas. IV. Functional characterization of ten diffuse histiocytic lymphoma cell lines. *Cancer* **42**, 2379–2391 (1978).
- Hogger, P., Dreier, J., Droste, A., Buck, F. & Sorg, C. Identification of the integral membrane protein RM3/1 on human monocytes as a glucocorticoid-inducible member of the scavenger receptor cysteine-rich family (CD163). *J. Immunol.* **161**, 1883–1890 (1998).
- Oliviero, S. & Cortese, R. The human haptoglobin gene promoter: interleukin-6-responsive elements interact with a DNA-binding protein induced by interleukin-6. *EMBO J.* **8**, 1145–1151 (1989).
- Mitani, K., Fujita, H., Kappas, A. & Sassa, S. Heme oxygenase is a positive acute-phase reactant in human Hep3B hepatoma cells. *Blood* **79**, 1255–1259 (1992).
- Andrews, N. C. & Levy, J. E. Iron is hot: an update on the pathophysiology of hemochromatosis. *Blood* **92**, 1845–1851 (1998).
- Donovan, A. et al. Positional cloning of zebrafish ferroportin1 identifies a conserved vertebrate iron exporter. *Nature* **403**, 776–781 (2000).
- Resnick, D., Pearson, A. & Krieger, M. The SRCR superfamily: a family reminiscent of the Ig superfamily. *Trends Biochem. Sci.* **19**, 5–8 (1994).
- Moestrup, S. K., Koltorf, K., Sottrup-Jensen, L. & Gliemann, J. The human α₂-macroglobulin receptor contains high affinity calcium binding sites important for receptor conformation and ligand recognition. *J. Biol. Chem.* **265**, 12623–12628 (1990).
- Moestrup, S. K. et al. β₂-glycoprotein-I (apolipoprotein H) and β₂-glycoprotein-I- phospholipid complex harbor a recognition site for the endocytic receptor megalin. *J. Clin. Invest.* **102**, 902–909 (1998).
- Birn, H. et al. Characterization of an epithelial approximately 460-kDa protein that facilitates endocytosis of intrinsic factor-vitamin B12 and binds receptor-associated protein. *J. Biol. Chem.* **272**, 26497–26504 (1997).
- Moestrup, S. K. & Gliemann, J. Analysis of ligand recognition by the purified α₂-macroglobulin receptor (low density lipoprotein receptor-related protein). Evidence that high affinity of α₂-macroglobulin-proteinase complex is achieved by binding to adjacent receptors. *J. Biol. Chem.* **266**, 14011–14017 (1991).
- Kozyraki, R. et al. The intrinsic factor-vitamin B12 receptor, cubilin, is a high-affinity apolipoprotein A-I receptor facilitating endocytosis of high-density lipoprotein. *Nature Med.* **5**, 656–661 (1999).
- Moestrup, S. K., Christensen, E. I., Sottrup-Jensen, L. & Gliemann, J. Binding and receptor-mediated endocytosis of pregnancy zone protein-proteinase complex in rat macrophages. *Biochim. Biophys. Acta* **930**, 297–303 (1987).

Acknowledgements

We thank K. Lassen and A.-M. Bundsgaard for technical assistance; K. Pulford for providing the SU-DHL-1 lymphoma cell line; and M. Sommerlund for inspiration.

Correspondence and requests for materials should be addressed to S.K.M. (e-mail: skm@biobase.dk). The gene sequence of the abundant variant of CD163 is deposited in GenBank under accession number Z22968.

Identification of the platelet ADP receptor targeted by antithrombotic drugs

Gunther Hollopeter^{*}, Hans-Michael Jantzen[†], Diana Vincent[†], Georgia Li[‡], Laura England[‡], Vanitha Ramakrishnan[†], Ruey-Bing Yang[†], Paquita Nurden[§], Alan Nurden[§], David Julius^{*} & Pamela B. Conley[†]

^{*} Department of Cellular and Molecular Pharmacology and Program in Neuroscience, University of California, San Francisco, San Francisco, California 94143, USA

[†] COR Therapeutics, Inc., South San Francisco, California 94080, USA

[§] UMR 5533 CNRS, Hôpital Cardiologique, 33605 Pessac, France

Platelets have a crucial role in the maintenance of normal haemostasis, and perturbations of this system can lead to pathological thrombus formation and vascular occlusion, resulting in stroke, myocardial infarction and unstable angina. ADP released from damaged vessels and red blood cells induces platelet aggregation through activation of the integrin GPIIb-IIIa and subsequent binding of fibrinogen. ADP is also secreted from platelets on activation, providing positive feedback that potentiates the actions of many platelet activators¹. ADP mediates platelet aggregation through its action on two G-protein-coupled receptor subtypes^{2,3}. The P2Y₁ receptor couples to G_q and mobilizes intracellular calcium ions to mediate platelet shape change and aggregation^{4,5}. The second ADP receptor required for aggregation (variously called P2Y_{ADP}, P2Y_{AC}, P2Y_{cyc} or P2T_{AC}) is coupled to the inhibition of adenylyl cyclase through G_i. The molecular identity of the G_i-linked receptor is still elusive, even though it is the target of efficacious antithrombotic agents, such as ticlopidine and clopidogrel^{6–8} and AR-C66096 (ref. 9). Here we describe the cloning of this receptor, designated P2Y₁₂, and provide evidence that a patient with a bleeding disorder¹⁰ has a defect in this gene. Cloning of the P2Y₁₂ receptor should facilitate the development of better antiplatelet agents to treat cardiovascular diseases.

To identify the G_i-linked platelet ADP receptor, we engineered *Xenopus* oocytes to allow the detection of G_i-linked responses through a sensitive electrophysiological assay. This strategy is based on the fact that several G_i-coupled receptors, such as the M2 muscarinic receptor, release Gβγ subunits from heterotrimeric G proteins, thereby activating inwardly rectifying K⁺ channels (Kir3.1–3.4)¹¹. A complementary DNA library from rat platelets was screened in oocytes expressing Kir3.1 and 3.4, and three positive pools that responded to 10 μM ADP (as determined by an increase in K⁺ current) were identified. Subfractionation of one of these pools led to the identification of a single clone tentatively designated as P2Y₁₂. The current induced by ADP was K⁺-dependent because replacement of K⁺ in the bath solution resulted in a complete loss of current (Fig. 1a). Additionally, injection of Kir or P2Y₁₂ *in vitro* RNA transcripts (cRNAs) alone gave no ADP-dependent currents, indicating that the observed signal was not due to activation of an endogenous purinergic receptor and was Kir-dependent (Fig. 1b). Moreover, when cRNA encoding pertussis toxin was injected together with the rat P2Y₁₂ clone, the response to ADP was abolished (Fig. 1b), as predicted for the G_i-linked platelet ADP receptor¹². The human P2Y₁₂ homologue was isolated from a human platelet library and similar results were obtained when this cRNA was expressed in *Xenopus* oocytes (Fig. 1a, b).

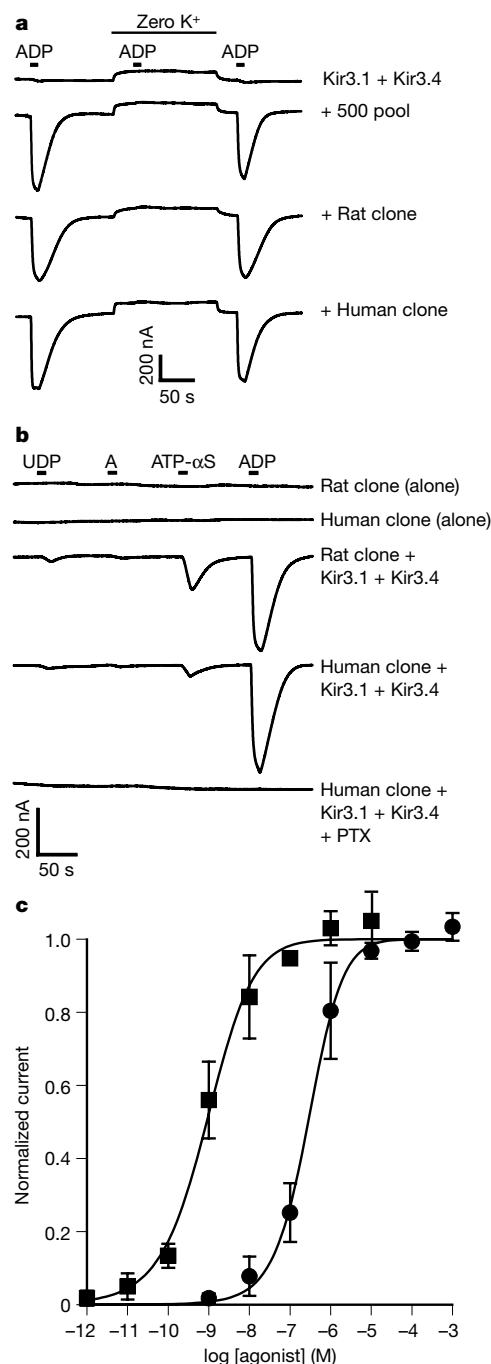


Figure 1 P2Y₁₂ is a G-protein-coupled receptor that responds to ADP. **a**, Activation of potassium-dependent currents in *Xenopus* oocytes expressing P2Y₁₂ with Kir3.1 and 3.4. ADP (10 μM) was applied (short bars) in the presence or absence (long bar) of extracellular potassium (70 mM) while recording membrane currents in the whole-cell voltage-clamp configuration. Oocytes injected with cRNA for Kir3.1 and 3.4 alone (top trace) do not exhibit significant currents in response to the application of ADP unless messages from a positive cDNA pool, the isolated rat P2Y₁₂ cRNA or the human P2Y₁₂ homologue are included (lower traces). **b**, ADP-selective stimulation of potassium channel-dependent currents by P2Y₁₂ occurs via a pertussis-toxin-sensitive pathway. UDP, adenosine (A), ATP-αS or ADP (10 μM each) were sequentially applied to oocytes expressing the rat or human receptor with or without Kir3.1, 3.4 and pertussis toxin (PTX). **c**, The agonist profile of P2Y₁₂ recapitulates that observed for the G_i-coupled platelet ADP receptor. Concentration–response curves for ADP (circles) and 2MeSADP (squares) are presented. Membrane currents were normalized in each oocyte to a response obtained with 10 μM ADP. Each point represents the mean (± s.d.) from five independent oocytes. The Hill equation was used to fit the response data.

[‡] Present addresses: Advanced Medicine, South San Francisco, California 94080, USA (G.L.); FibroGen, South San Francisco, California 94080, USA (L.E.).

One hallmark of the G_i -linked platelet ADP receptor is that substitution of alkylthio groups at the 2-position of the adenine ring increases potency at the receptor^{1,13,14}. Consistent with this, 2-methylthioadenosine 5'-diphosphate (2MeSADP) displayed two orders of magnitude more potency than ADP (with half-maximal responses at 0.9 and 300 nM, respectively) (Fig. 1c). In contrast,

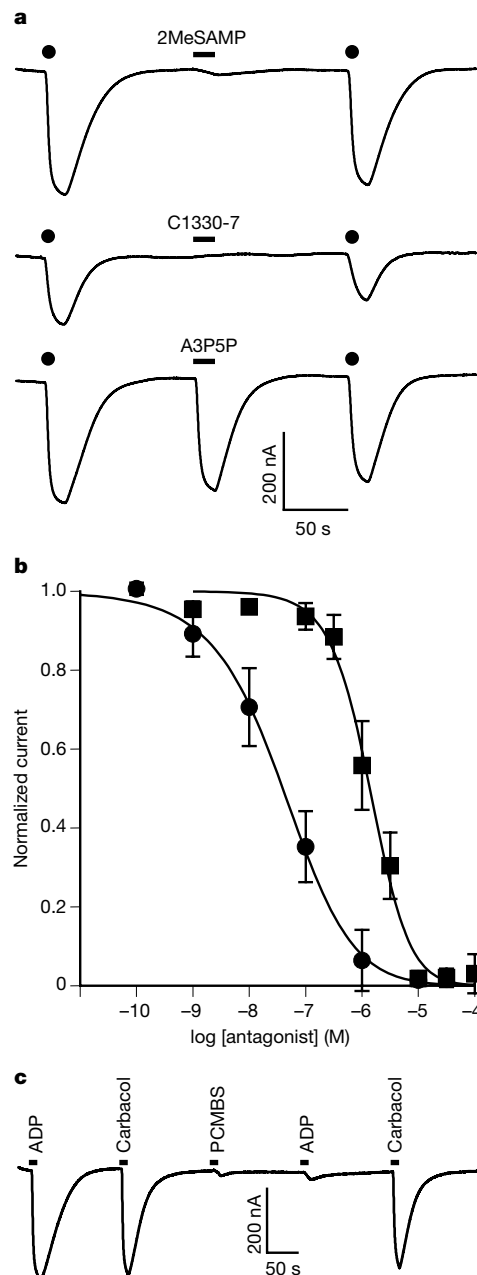


Figure 2 Currents stimulated by ADP in oocytes expressing hP2Y₁₂ with Kir3.1 and 3.4 are inhibited by 2MeSAMP, C1330-7 or a thiol reagent. **a**, Current tracing showing reversible blocking of ADP (1 μ M) responses by 2MeSAMP (10 μ M) or C1330-7 (1 μ M), but not by A3P5P (300 μ M). The dots indicate the start of 15-s applications of ADP; bars indicate co-application with the antagonist. **b**, 2MeSAMP (squares) and C1330-7 (circles) inhibition curves. Current responses were normalized to that elicited by ADP (500 nM) alone in each oocyte and plotted as means $\pm$ s.d. Curves were fitted to the data with the Hill equation ($n = 5$ independent oocytes for each point). **c**, Selective ablation of P2Y₁₂ but not M2 muscarinic receptor signalling by the thiol reagent pCMBS. ADP (10 μ M), carbachol (1 μ M) and pCMBS (1 mM) were applied sequentially to an oocyte expressing both receptors concurrently with Kir3.1 and 3.4. Bars indicate periods of drug application (10 s).

other nucleoside or nucleotide derivatives had no effect (Fig. 1b). We also examined the actions of several antagonists specific for the platelet G_i -linked ADP receptor. Treatment of *Xenopus* oocytes expressing the rat or human P2Y₁₂ receptor with the nucleotide derivative 2-methylthioadenosine 5'-monophosphate (2MeSAMP)² or a non-nucleotide inhibitor C1330-7 (ref. 15) blocked ADP-induced K^+ currents with half-maximal inhibitory concentrations of 1.4 μ M and 40 nM, respectively (Fig. 2b). In contrast, the P2Y₁-selective antagonist A3P5P (ref. 16) had no inhibitory effect on the signal evoked by ADP at the rat or human P2Y₁₂ (Fig. 2a). Thus, when expressed in *Xenopus* oocytes, the P2Y₁₂ receptor recapitulates the pharmacological profile previously described for the platelet G_i -linked ADP receptor. The only anomaly that we observed relates to the action of ATP- α S, which behaved as a weak agonist rather than as an antagonist at the cloned receptor. This finding was somewhat unexpected because ATP derivatives reportedly antagonize the platelet G_i -linked receptor. However, this discrepancy might reflect partial degradation or impurities in commercially available preparations of ATP- α S, or differences between the platelet and oocyte environments such as the degree of ectonucleotidase activity. Indeed, recombinant P2Y₁ receptors respond differentially to ATP, depending on the expression system used^{17,18}.

Chinese hamster ovary (CHO) cells expressing the hP2Y₁₂ receptor displayed ADP-mediated repression of forskolin-stimulated cyclic AMP (cAMP) levels in a dose-dependent manner, reaching a maximum of 47% reduction at 10 μ M ADP (Fig. 3a). The

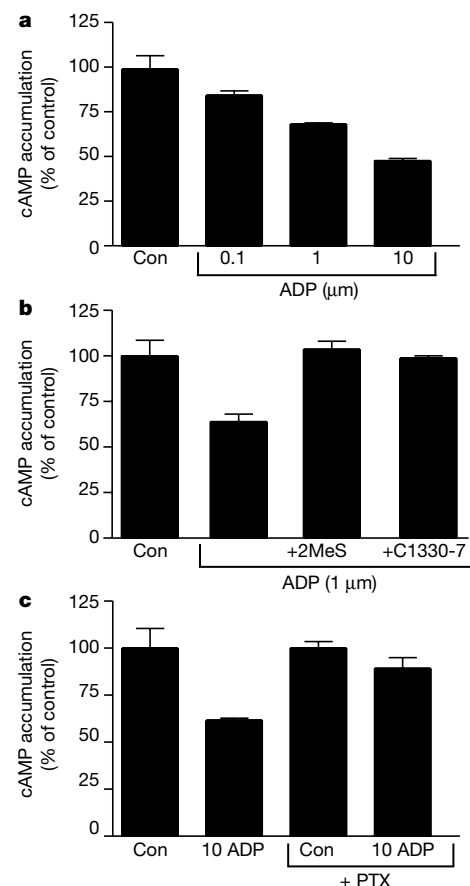


Figure 3 Activation of hP2Y₁₂ in CHO cells inhibits adenyl cyclase. **a**, Receptor coupling to adenyl cyclase was assessed as ADP-mediated (0.1–10 μ M) inhibition of forskolin-stimulated (10 μ M) cAMP accumulation (Con, control; normalized to 100%). **b**, Effect of the specific antagonists 2MeSAMP (2MeS) (50 μ M) and C1330-7 (50 μ M) on repression of ADP-mediated (1 μ M) forskolin-stimulated cAMP levels. **c**, Effect of pretreatment with PTX on the inhibition by 10 μ M ADP of forskolin-stimulated cAMP levels. Results are means $\pm$ s.d. of three representative experiments performed in triplicate.

repression of cAMP levels by 1 μ M ADP was reversed by the selective antagonists 2MeSAMP and C1330-7 (Fig. 3b), in agreement with the pharmacological profile observed in *Xenopus* oocytes, and as described for the G_i -coupled receptor on platelets. Neither of these antagonists had effects on forskolin-stimulated cAMP levels in the absence of agonist. Similar responses to ADP were observed in rat 2-9 fibroblasts stably expressing rP2Y₁₂ (data not shown). Pretreatment of transfected cells with pertussis toxin abolished effects of ADP on forskolin-stimulated cAMP (Fig. 3c), indicating

that the response is G_i -mediated.

Northern blot analysis demonstrated that P2Y₁₂ is abundantly expressed in human platelets, and to a smaller extent in brain (Fig. 4a, b). The predominant transcript of 2.4 kb was absent from all other tissues examined, including peripheral-blood leukocytes. A fainter species of ~4.5 kb was also detected in platelet and brain, whereas a prominent band of ~1.0 kb (Fig. 4b) was observed only in platelet RNA. We also found selective expression in platelets and brain in rat tissues (data not shown). Thus, the mRNA for this novel G-protein-coupled receptor (GPCR) has a restricted expression pattern and is abundantly present in platelets, which is consistent with this cDNA encoding the platelet G_i -linked receptor. Within the brain, the 2.4-kb species was observed in numerous subregions, including the amygdala, caudate nucleus, corpus callosum, hippocampus, substantia nigra and thalamus (data not shown). Cellular resolution of the expression of rP2Y₁₂ was obtained by *in situ* hybridization histochemistry of brain sections; punctate staining was noted throughout white and grey matter (Fig. 4c). Principal cells of the hippocampus were not stained, nor was a laminar pattern of expression observed in the neocortex. These observations are consistent with a glial expression pattern. Interestingly, the only cell line previously described to express a P2Y purinergic receptor that is negatively coupled to adenylyl cyclase is the rat C6 glioma cell line¹⁹. Indeed, we detected a 2.4-kb mRNA species in these cells by northern blot analysis with a rP2Y₁₂ probe (data not shown).

By using a rabbit polyclonal antiserum directed against the predicted amino terminus of rP2Y₁₂, we assessed the surface expression of receptor protein on stably transfected rat 2-9 fibroblasts or rat platelets by flow cytometry. At an antibody concentration of 25 μ g ml⁻¹, ninefold (Fig. 4e) and fourfold (Fig. 4d) increases in mean fluorescence intensity (compared with a control antibody) were observed with transfected cells and platelets, respectively, demonstrating that P2Y₁₂ protein is indeed expressed on the platelet surface.

When the chromosomal localization of the P2Y₁₂ gene was determined with the Stanford G3 panel²⁰ (Research Genetics), P2Y₁₂-specific primers mapped closest to STS-D13626, corresponding to the KIAA0001 gene recently identified as a UDP-glucose GPCR²¹. Both of these genes reside on chromosome 3q24–25, interval D3S1279–1280, a region that also includes the human P2Y₁ gene, (GeneMap 99; <http://www.ncbi.nlm.nih.gov>). Thus, this interval contains genes encoding at least three receptors, two of which (P2Y₁ and P2Y₁₂) mediate ADP-dependent platelet aggregation. Among GPCRs, P2Y₁₂ is most closely related to the UDP-glucose receptor²¹ (44% identical) but much less so to P2Y₁ (19% identical), suggesting that the UDP-glucose and P2Y₁₂ receptors are the products of a relatively recent gene duplication on chromosome 3.

The predicted hP2Y₁₂ protein contains four extracellular cysteine residues (see Fig. 5). A critical role of cysteine residues in the function of the platelet ADP receptor has been suggested by the ability of thiol reagents to ablate ADP responses in platelets¹. Indeed, the antithrombotic agent clopidogrel is proposed to inactivate the G_i -linked platelet ADP receptor through a mechanism in which it is metabolized to a thiol species that modifies a cysteine residue on the receptor²². We found that a brief exposure of oocytes expressing Kir3.1, 3.4 and hP2Y₁₂ to the thiol reagent *p*-chloromercuriphenylsulphonic acid (pCMBS) eliminated ADP-evoked current responses (Fig. 2c). Inhibition was selective for the P2Y₁₂ receptor because the activation of this signalling pathway by M2 muscarinic receptors expressed in the same oocytes was unaffected by treatment with pCMBS.

Nurden *et al.*¹⁰ have previously described a patient (M.L.) with a mild bleeding disorder. Platelets from M.L. exhibit impaired ADP-dependent platelet aggregation, greatly reduced ADP binding activity and lack the ability to inhibit cAMP levels in response to ADP. However, the P2Y₁-receptor-mediated responses, such as

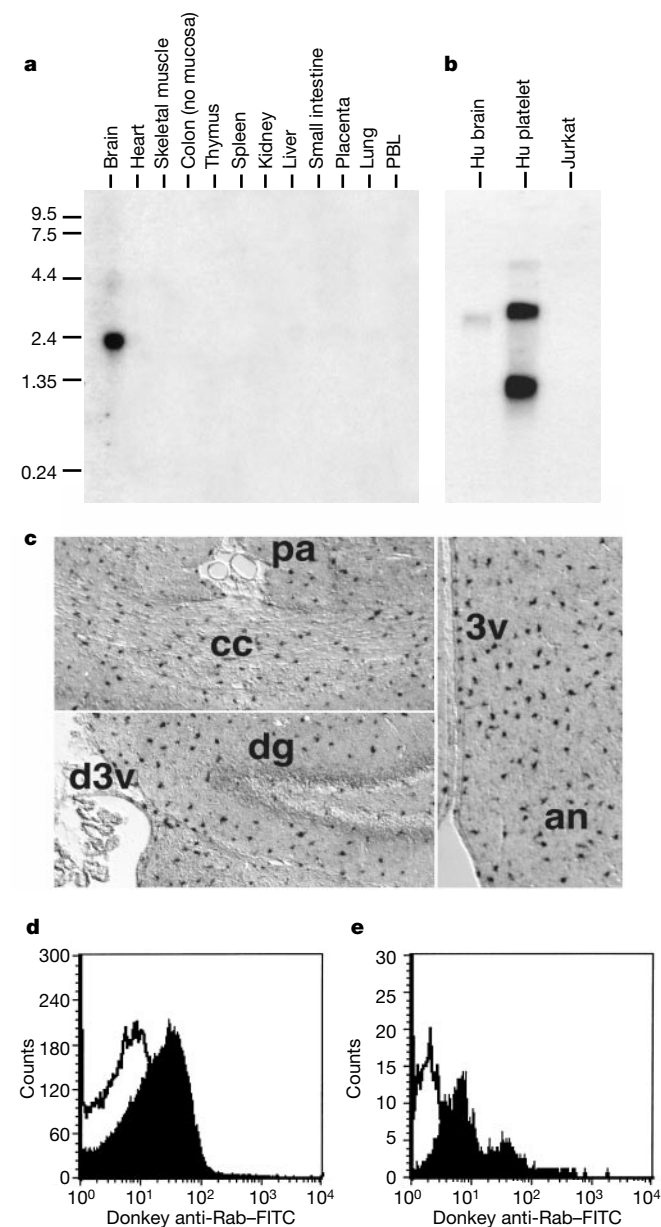


Figure 4 P2Y₁₂ receptor is selectively expressed in platelets and brain. **a, b**, Northern analysis of hP2Y₁₂ transcripts. All lanes contained 2 μ g poly(A)⁺ mRNA except samples from platelet and Jurkat cells, which contain 20 μ g total RNA. PBL, peripheral blood lymphocytes; Hu, human. Kilobases (Kb) are indicated on the left. **c**, rP2Y₁₂ transcripts are distributed throughout the brain in presumptive glia. Staining was equally abundant in fibre tracts (cc, corpus callosum) and regions enriched for neuronal cell bodies (dg, dentate gyrus; an, arcuate nucleus of the hypothalamus), but was absent from vasculature (pa, pericallosal artery). Control (sense) riboprobes did not stain these regions. Ventricular structures are also indicated (d3v, dorsal third ventricle; 3v, third ventricle). **d**, FACS analysis of rat platelets stained with rP2Y₁₂ antisera (filled peak) or a control IgG (unfilled peak). Rab, rabbit. **e**, FACS analysis of rat 2-9 fibroblasts transfected with the rP2Y₁₂ cDNA clone (filled peak) or untransfected rat 2-9 fibroblasts (open peak).

intracellular Ca^{2+} ion mobilization and shape change, are not affected, indicating that this patient has a selective defect in the G_i -linked receptor. Analysis of polymerase chain reaction (PCR) products from the P2Y_{12} coding region from M.L.'s genomic DNA revealed the presence of one mutant allele at this locus, as confirmed by direct sequencing of at least three independent PCR reactions. The mutation found in the gene encoding P2Y_{12} consists of a deletion of two nucleotides (TTCATT) within the coding region, at amino acid 240 (near the N-terminal end of TM6), thus shifting the reading frame for 28 residues before introducing a premature stop codon (Fig. 5b). This mutation was not detected in 100 randomly chosen, ethnically matched control individuals. Biochemical studies suggest that platelets from M.L. lack G_i -linked ADP receptors, yet our sequence analysis indicates that this

individual has one mutant and one wild-type P2Y_{12} allele, at least in the protein-coding region. This suggests either that the P2Y_{12} mutation identified exerts a dominant-negative effect or that M.L. harbours an additional mutation that eliminates the expression of the allele containing a wild-type coding region. We evaluated the former possibility with our electrophysiological assay (Fig. 5c). First, no significant activity was observed when oocytes were injected with cRNA transcripts corresponding to the frame-shifted allele, demonstrating that this mutant is indeed non-functional. Moreover, when mutant and wild-type cRNAs were injected together into oocytes at different ratios, no inhibition of the signal from the wild-type allele was observed, demonstrating that the mutant allele does not act in a dominant-negative manner. Further support for this conclusion comes from sequence analysis of the P2Y_{12} coding region from

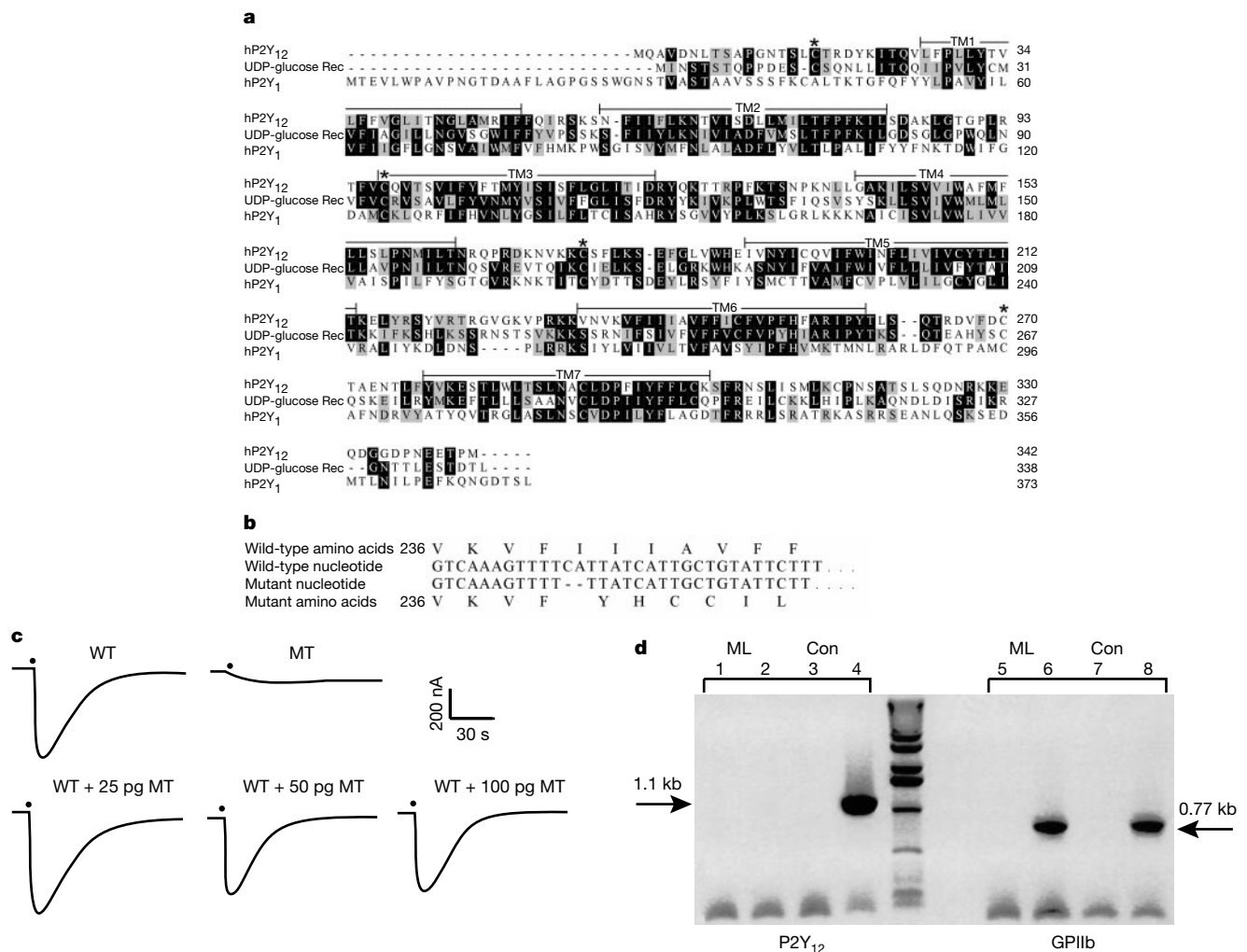


Figure 5 A frame-shift mutation within the hP2Y_{12} gene is associated with a bleeding disorder. **a**, Deduced amino acid sequence of the hP2Y_{12} protein and its alignment with other homologous receptor sequences. The putative membrane-spanning domains are designated with bars above the sequence. hP2Y_{12} sequence is aligned with the sequences of hP2Y_1 receptor (also expressed in platelets and activated by ADP), as well as with the human UDP-glucose receptor, with which it shares the greatest homology. Shading denotes amino acid identity (black) or similarity (grey); asterisks denote extracellular cysteine residues. **b**, A P2Y_{12} allele from a patient (M.L.) with defective ADP-dependent aggregation contains a 2-base-pair deletion, resulting in a frame-shift mutation and a premature truncation of the protein. **c**, Mutant hP2Y_{12} receptor from patient M.L. is non-functional and does not act in a dominant-negative capacity. Representative ADP-evoked membrane currents from an oocyte injected with 50 pg wild-type (WT) hP2Y_{12} cRNA (upper left panel), 50 pg mutant (MT) hP2Y_{12} cRNA (upper right

panel) or with 50 pg WT and increasing amounts of MT hP2Y_{12} cRNAs (bottom panels). Oocytes were also injected with 1 ng Kir3.1 and 1 ng Kir3.4 cRNAs. Dots indicate the onset of ADP application ($10 \mu\text{M}$ for 5 s). **d**, Patient M.L. has abnormally low levels of RT-PCR product derived from P2Y_{12} mRNA. RT-PCR with primers specific for either P2Y_{12} (lanes 1–4) or GPIIb (lanes 5–8) was performed with whole-blood RNA from patient M.L. (lanes 1, 2, 5 and 6) or a control (Con) sample (lanes 3, 4, 7 and 8). PCR reactions were performed on RNA samples without reverse transcriptase to control for genomic DNA contamination (lanes 1, 3, 5 and 7). A 1.1-kb product encoding the P2Y_{12} open reading frame was amplified from the control sample but was virtually absent from M.L. (a faint product could be observed after longer exposure). In contrast, the amount of product (0.77 kb) amplified from GPIIb mRNA was equivalent between M.L. and control. Sequence analysis reveals that M.L.'s P2Y_{12} RT-PCR product was derived solely from the mutant allele.

M.L.'s daughter, who has previously been shown to have an intermediate number of ADP-binding sites and impaired ADP-dependent aggregation¹⁰. Like her father, she has one wild-type and one frame-shifted allele, and is therefore likely to be a true heterozygote, both phenotypically and genotypically. If so, then the truncated receptor does not act as a dominant-negative *in vivo*. Finally, we asked whether M.L.'s alleles are both expressed by performing a PCR analysis after reverse transcription (RT-PCR) of RNA from his platelets. Extremely low levels of P2Y₁₂-derived product were obtained compared with levels amplified from an unaffected individual or compared with a control transcript encoding platelet GPIIb (Fig. 5D). In addition, a sequence analysis of P2Y₁₂ RT-PCR products demonstrated that M.L.'s P2Y₁₂ transcripts are derived from the mutant allele; no wild-type product was detected. We therefore conclude that M.L.'s lack of functional G_i-coupled platelet ADP receptor activity occurs because he expresses only the frame-shifted allele. The mechanism by which expression of the wild-type allele is repressed is not clear. Other patients with congenital platelet defects have been reported²³, and it will be interesting to determine whether these are also associated with mutations in P2Y₁₂.

We have characterized a novel cDNA from a platelet library that encodes the G_i-linked platelet ADP receptor. Genetic^{4,5,10,23} and pharmacological^{24,25} studies demonstrate that the G_i-linked receptor is critical for the formation and stabilization of large platelet aggregates²⁶. Additionally, the G_i-linked receptor is the target of the antithrombotic drugs clopidogrel and ticlopidine, which have been demonstrated to be effective in treating a variety of thrombotic diseases (stroke, myocardial infarction, peripheral vascular disease). However, these drugs work through a mechanism of covalent protein modification, which might underlie their recent association with the syndrome thrombotic thrombocytopenic purpura²⁷, an immune-mediated response. Our studies demonstrate that the P2Y₁₂ receptor has a selective tissue distribution compared with other purinergic receptors (such as P2Y₁), making this receptor an extremely attractive target for the development of new anti-thrombotic drugs. □

Methods

Platelet cDNA library

Poly-A⁺ mRNA from rat platelets was used to generate a directional oligo(dT)-primed cDNA library in the pCDNA3.1⁺ vector. Roughly 320,000 clones were divided into 48 individual pools. Linearized cDNA templates from these pools were transcribed *in vitro* with T7 RNA polymerase (Ambion). Sib selection of a positive pool was performed to subfractionate the signal to the level of 96 clones. All were sequenced and a novel GPCR was further characterized. Rat P2Y₁₂ cDNA (GenBank accession number AF313450) was used to isolate a human orthologue from a platelet λ ZAP cDNA library. A full-length hP2Y₁₂ cDNA expression construct was obtained by ligation of a λ-clone and a fragment derived by 3' rapid amplification of cloned ends into the pcNeo expression vector (Promega). The GenBank accession number for human P2Y₁₂ is AF313449.

Platelet RT-PCR

Informed consent was obtained from M.L. and his daughter for the purposes of this study. Whole blood (30 ml) was lysed and total RNA was isolated with TriReagent BD (Molecular Research Center). First-strand cDNA was generated (Superscript 2, Life Technologies) and PCR (35 cycles) was performed with the following mRNA-specific primers. The P2Y₁₂ primers 5' (5'-CCAGAATCAACAGTTATCAGGTAACC-3') and 3' (5'-GTCTGCTTATATTTTACTTAGCGCTTTGC-3') were annealed at 57°C, whereas the GPIIb primers 5' (5'-GTCAACGGGGATGGGAGGCATGA-3') and 3' (5'-GTCTGCCTCATCTCGAAGGAAGG-3') were annealed at 60°C. PCR products were analysed by electrophoresis in 1% agarose, and bands of the correct size were isolated for direct sequencing.

Electrophysiology

Defolliculated *Xenopus laevis* oocytes were injected with a positive 500 clone pool (10 ng), rP2Y₁₂ (10 pg), hP2Y₁₂ (50 pg), Kir3.1, Kir3.4, PTX or hM2 (1 ng each) cRNAs as indicated. Three to seven days after injection, two-electrode voltage-clamp recordings were performed with a Geneclamp 500 amplifier (Axon Instruments) and a MacLab analogue-to-digital converter (MacLab). Membrane potentials were clamped at -70 mV while the recording chamber was perfused at a rate of 2 ml min⁻¹ with a solution containing (in mM) 70 KCl, 20 NaCl, 3 MgCl₂, 5 HEPES, pH 7.4, at room temperature. The KCl was replaced with NaCl to examine responses in the absence of potassium. Agonists and antagonists (Roche Molecular Biochemistry or Sigma) were diluted in the

recording solution. Experiments with C1330-7 included 0.1% dimethyl sulphoxide to enhance its solubility in the perfusate.

Generation of stable mammalian cell lines and cAMP assays

Chinese hamster ovary cells or rat 2-9 fibroblasts, which are null for G_i-linked purinergic receptors, were transfected with hP2Y₁₂ or rP2Y₂ cDNAs, respectively, with FuGene reagent (Roche), and cells were cultured in the presence of G418 for 2 weeks to select for stable transfectants. For cAMP assays, stably transfected CHO cells expressing the hP2Y₁₂ plasmid were plated in 12-well dishes. After 48 h, medium was removed from the cells and replaced with serum-free medium containing IBMX (0.25 mM final) and incubated at 37°C for 5 min. Cells were incubated for an additional 5 min with 10 μM forskolin, as well as the indicated agonists and antagonists. Pertussis toxin treatment (30 ng/ml) occurred for 20 h at 37°C before assay. Levels of cAMP were determined from aliquots of cell extracts in a radioimmunoassay (Amersham Biotrak cAMP¹²⁵I assay system).

Northern and *in situ* hybridizations

Northern blots of poly(A)⁺ RNA from human tissues (Clontech) or total human platelet RNA was hybridized with radiolabelled hP2Y₁₂ cDNA fragments under standard conditions. Digoxigenin-labelled *in situ* hybridization was performed on coronal rat brain sections by using an RNA probe corresponding to the antisense sequence of rP2Y₁₂ (ref. 28).

Flow cytometry

Adult male Sprague-Dawley rats were anaesthetized and whole blood was isolated; citrate was used as anticoagulant. Platelet-rich plasma was isolated by centrifugation and used for flow-cytometry analysis. A rabbit anti-serum (SynPep Corporation) was produced to the N-terminal 23 residues of rP2Y₁₂. IgG was purified with protein G-Sepharose. Rat platelet-rich plasma (2 × 10⁶ cells) and cultured rat 2-9 fibroblasts transfected with rP2Y₁₂ cDNA (10⁵ cells) were incubated with purified IgG (10–50 μg/ml) in FACS buffer (phosphate-buffered saline containing 0.1% BSA and 2% heat-inactivated fetal bovine serum) in a total volume of 100 μl for 1 h at 4°C. Cells and platelets were then washed with cold FACS buffer and incubated with 2.5 μg ml⁻¹ FITC-conjugated goat anti-rabbit antibody for 30 min at 4°C. Cells and platelets were washed, resuspended in cold FACS buffer, and fluorescence of cell-bound secondary antibody was determined with a FACSort flow cytometer (Becton-Dickinson). Control samples contained cells without antibodies (for the determination of autofluorescence), cells with control rabbit IgG, or secondary antibodies alone.

Received 13 June; accepted 10 October 2000.

- Mills, D. C. ADP receptors on platelets. *Thromb. Haemost.* **76**, 835–856 (1996).
- Jantzen, H. M. *et al.* Evidence for two distinct G-protein-coupled ADP receptors mediating platelet activation. *Thromb. Haemost.* **81**, 111–117 (1999).
- Daniel, J. L. *et al.* Molecular basis for ADP-induced platelet activation. I. Evidence for three distinct ADP receptors on human platelets. *J. Biol. Chem.* **273**, 2024–2029 (1998).
- Leon, C. *et al.* Defective platelet aggregation and increased resistance to thrombosis in purinergic P2Y₁ receptor-null mice. *J. Clin. Invest.* **104**, 1731–1737 (1999).
- Fabre, J. E. *et al.* Decreased platelet aggregation, increased bleeding time and resistance to thromboembolism in P2Y₁-deficient mice. *Nature Med.* **5**, 1199–1202 (1999).
- Gachet, C. *et al.* The thienopyridine ticlopidine selectively prevents the inhibitory effects of ADP but not of adrenaline on cAMP levels raised by stimulation of the adenylate cyclase of human platelets by PGE₁. *Biochem. Pharmacol.* **40**, 2683–2687 (1990).
- Gachet, C. *et al.* ADP receptor induced activation of guanine nucleotide binding proteins in rat platelet membranes—an effect selectively blocked by the thienopyridine clopidogrel. *Thromb. Haemost.* **68**, 79–83 (1992).
- Mills, D. C. B. *et al.* Clopidogrel inhibits the binding of ADP analogues to the receptor mediating inhibition of platelet adenylate cyclase. *Arterioscler. Thromb.* **12**, 430–436 (1992).
- Humphries, R. G., Tomlinson, W., Ingall, A. H., Cage, P. A. & Leff, P. A novel, highly potent and selective antagonist at human platelet P_{2U}-purinoreceptors. *Br. J. Pharmacol.* **113**, 1057–1063 (1994).
- Nurden, P. *et al.* An inherited bleeding disorder linked to a defective interaction between ADP and its receptor on platelets. *J. Clin. Invest.* **95**, 1612–1622 (1995).
- Krapivinsky, G., Krapivinsky, L., Wickman, K. & Clapham, D. Gβγ binds directly to the G protein-gated K⁺ channel, IKACH. *J. Biol. Chem.* **270**, 29059–29062 (1995).
- Ohlmann, P. *et al.* The human platelet ADP receptor activates G₁₂ proteins. *Biochem. J.* **312**, 775–779 (1995).
- MacFarlane, D. E., Srivastava, P. C. & Mills, D. C. B. 2-Methylthioadenosine[β-³²P]diphosphate: an agonist and radioligand for the receptor that inhibits the accumulation of cyclic AMP in intact blood platelets. *J. Clin. Invest.* **71**, 420–428 (1983).
- Hourani, S. M. O. & Hall, D. Receptors for ADP on human blood platelets. *Trends Pharmacol. Sci.* **15**, 103–108 (1994).
- Jantzen, H.-M. *et al.* Evidence for two distinct G protein-coupled ADP receptors mediating platelet activation. *Blood* **92**, 303a (1998).
- Boyer, J. L., Romero-Avila, T., Schachter, J. B. & Harden, T. K. Identification of competitive antagonists of the P2Y₁ receptor. *Mol. Pharmacol.* **50**, 1323–1329 (1996).
- Palmer, R. K., Boyer, J. L., Schachter, J. B., Nicholas, R. A. & Harden, T. K. Agonist action of adenosine triphosphates at the human P2Y₁ receptor. *Mol. Pharmacol.* **54**, 1118–1123 (1998).
- Filippov, A. K., Brown, D. A. & Barnard, E. A. The P2Y₁ receptor closes the N-type Ca²⁺ channel in neurones, with both adenosine triphosphates and diphosphates as potent agonists. *Br. J. Pharmacol.* **129**, 1063–1066 (2000).
- Boyer, J. L., Lazarowski, E. R., Chen, X. H. & Harden, T. K. Identification of a P2Y₁-purinergic receptor that inhibits adenylate cyclase. *J. Pharmacol. Exp. Ther.* **267**, 1140–1146 (1993).
- Stewart, E. *et al.* An STS-based radiation hybrid map of the human genome. *Genome Res.* **7**, 422–433 (1997).

21. Chambers, J. K. *et al.* A G protein-coupled receptor for UDP-glucose. *J. Biol. Chem.* **275**, 10767–10771 (2000).
22. Savi, P. *et al.* Identification and biological activity of the active metabolite of Clopidogrel. *Thromb. Haemost.* (in the press)
23. Cattaneo, M. & Gachet, C. ADP receptors and clinical bleeding disorders. *Arterioscler. Thromb. Vasc. Biol.* **19**, 2281–2285 (1999).
24. Jarvis, G. E., Humphries, R. G., Robertson, M. J. & Leff, P. ADP can induce aggregation of human platelets via both P2Y₁(1) and P2T) receptors. *Br. J. Pharmacol.* **129**, 275–282 (2000).
25. Hechler, B., Eckly, A., Ohlmann, P., Cazenave, J.-P. & Gachet, C. The P2Y₁ receptor, necessary but not sufficient to support full ADP-induced platelet aggregation, is not the target of the drug clopidogrel. *Br. J. Haematol.* **103**, 858–866 (1998).
26. Humbert, M. *et al.* Ultrastructural studies of platelet aggregates from human subjects receiving clopidogrel and from a patient with an inherited defect of an ADP-dependent pathway of platelet activation. *Arterioscler. Thromb. Vasc. Biol.* **16**, 1532–1543 (1996).
27. Bennett, C. L. *et al.* Thrombotic thrombocytopenic purpura associated with clopidogrel. *N. Engl. J. Med.* **325**, 1371–1372 (2000).
28. Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).

Acknowledgements

We thank E. Peralta for advice and encouragement. We also thank H. Chuang for advice with electrophysiological assays; A. Laibelman and R. Scarborough for C1330-7; P. Castro and L. Komuves for assistance and advice with the *in situ* hybridization; the Oksenberg and Reijo laboratories for human DNA samples; E. Thompson, K. Simpson, S. Hollenbach and K. Ministri-Madrid and the COR animal facility for technical assistance; R. Wong and D. De Guzman for assistance with figures; and C. Homcy, E. E. Reynolds, J. Topper and D. Phillips for discussions. D.J. is supported by funding from NIH and NIMH; G.H. is supported by a NIH predoctoral neuroscience training grant.

Correspondence and requests for materials should be addressed to P.B.C. (e-mail: pconley@corr.com).

Inactivation of the apoptosis effector *Apaf-1* in malignant melanoma

María S. Soengas*, **Paola Capodice†**, **David Polsky†**, **Jaume Mora†**, **Manel Esteller‡**, **Ximena Opitz-Araya***, **Richard McCombie***, **James G. Herman‡**, **William L. Gerald†**, **Yuri A. Lazebnik***, **Carlos Cordon-Cardó†** & **Scott W. Lowe***

* Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

† Memorial Sloan Kettering Cancer Center, New York, New York 10021, USA

‡ Johns Hopkins Oncology Center and Johns Hopkins University, Baltimore, Maryland 21231, USA

Metastatic melanoma is a deadly cancer that fails to respond to conventional chemotherapy and is poorly understood at the molecular level¹. *p53* mutations often occur in aggressive and chemoresistant cancers but are rarely observed in melanoma^{1,2}. Here we show that metastatic melanomas often lose *Apaf-1*, a cell-death effector that acts with cytochrome *c* and caspase-9 to mediate *p53*-dependent apoptosis³. Loss of *Apaf-1* expression is accompanied by allelic loss in metastatic melanomas, but can be recovered in melanoma cell lines by treatment with the methylation inhibitor 5-aza-2'-deoxycytidine (5aza2dC). *Apaf-1*-negative melanomas are invariably chemoresistant and are unable to execute a typical apoptotic programme in response to *p53* activation. Restoring physiological levels of *Apaf-1* through gene transfer or 5aza2dC treatment markedly enhances chemosensitivity and rescues the apoptotic defects associated with *Apaf-1* loss. We conclude that *Apaf-1* is inactivated in metastatic melanomas, which leads to defects in the execution of apoptotic cell death. *Apaf-1* loss may contribute to the low frequency of *p53* mutations observed in this highly chemoresistant tumour type.

Mutations in the *p53* tumour-suppressor gene are extremely common in most human cancers, and are often associated with poor patient prognosis and treatment failure⁴. *p53* promotes cell-cycle arrest or apoptosis in response to cellular damage, including

that induced by many chemotherapeutic drugs⁴. As a consequence, *p53* loss can promote drug resistance in experimental and clinical settings^{4,5}. Caspase-9 (Casp9) and its cofactor *Apaf-1* can be essential downstream effectors of *p53* during apoptosis, such that inactivation of either *Apaf-1* or *Casp9* substitutes for the loss of *p53* in enhancing the transformation potential of primary murine fibroblasts³, and in promoting chemoresistance^{3,6,7}. Hence, *Apaf-1* and *Casp9* are potential tumour suppressors that may affect treatment sensitivity.

To investigate whether *Apaf-1* is a tumour suppressor, we surveyed several tumour types for loss-of-function mutations in the *Apaf-1* gene. We included melanoma in this analysis because it is extremely resistant to apoptosis-inducing agents but retains wild-type *p53* (ref. 1). We analysed a series of metastatic melanoma samples (see Supplementary Information) for their *p53* and *Apaf-1* status. As expected, a low rate of mutation was detected for *p53* in these tumours (see Supplementary Information). In contrast, a markedly high rate of allelic loss was found for polymorphic markers encompassing the *Apaf-1* locus on chromosome 12q23 (> 40%; Fig. 1a and b). Moreover, tumours with loss of heterozygosity (LOH) expressed little *Apaf-1* message by *in situ* hybridization (Fig. 1c, d), compared with tumours with no LOH (Fig. 1c, d).

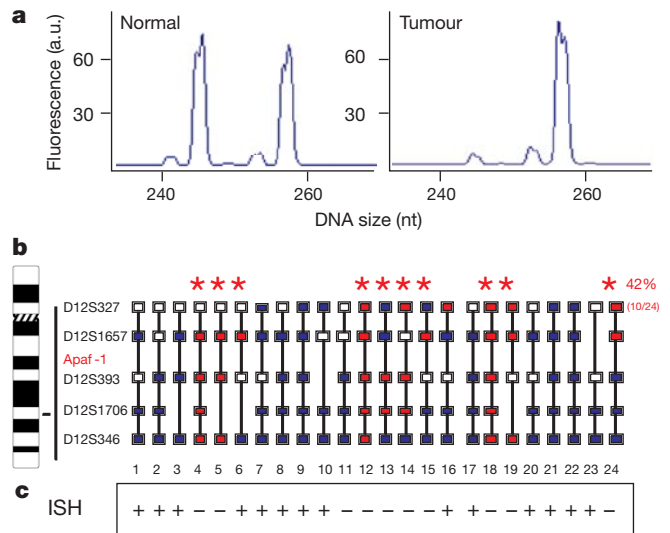


Figure 1 *Apaf-1* loss in metastatic melanoma. **a**, Fluorescent electropherograms for microsatellite marker D12S393 on chromosome 12q22–23. The normal (left) and tumour (right) counterparts of case four are indicated. a.u., arbitrary units. **b**, Summary of LOH analysis for 24 normal/tumour pairs, showing non-informative (homozygous) markers (white), retained markers (blue) and markers with LOH (red). Each vertical line indicates a single patient, and cases having LOH at *Apaf-1* flanking markers are indicated with a red asterisk. **c**, Estimation of *Apaf-1* messenger RNA levels by *in situ* hybridization using *Apaf-1* antisense probe (residues 2,112–2,225). Relative levels are indicated as high (+) and low or undetectable (–). **d**, Representative examples of samples described in **c**. Left, tumour case 7 (no LOH); right, tumour case 19 (LOH). A 'sense' probe control produced no signal (data not shown). Similar results were obtained for probes targeting other *Apaf-1* regions (residues 1–461, 581–1,363 or 3,426–3,710) (data not shown).

Only one out of six primary melanomas analysed showed reduced Apaf-1 expression (see Supplementary Information). Thus, loss of Apaf-1 is a relatively common event in melanoma that may be associated with progression of the disease.

We also examined *Apaf-1* expression in a series of 19 cell lines derived from metastatic melanomas (see Supplementary Information). As determined by protein immunoblotting, five lines expressed Apaf-1 levels comparable to those of normal melanocytes ('Apaf-1 positive'); four lines expressed intermediate Apaf-1 levels (70–30%); and ten lines expressed little Apaf-1 ('Apaf-1 negative') (Fig. 2a; see Supplementary Information). Similarly, northern blotting (Fig. 2b) and *in situ* hybridization (Fig. 2c) showed that Apaf-1-negative cells expressed low or undetectable levels of *Apaf-1* mRNA. Analysis of genomic DNA using fluorescence *in situ* hybridization and TaqMan quantitative polymerase chain reaction (PCR) (Supplementary Information; and data not shown) indicated that several Apaf-1-negative lines harboured only a single copy of the *Apaf-1* gene, although none showed homozygous deletions (Supplementary Information; data not shown). Moreover, mutations in *Apaf-1* were not detected in eight of the cell lines that we tested (3, 5–11; data not shown). These results suggest that the remaining *Apaf-1* allele is transcriptionally repressed.

Gene silencing that arises from methylation is an important epigenetic mechanism of tumour-suppressor inactivation^{8,9}. Methylated CpGs are recognized by proteins that recruit histone deacetylases, leading to stable transcriptional repression, which can often be reversed by the methylation inhibitor 5-aza2dC or by the histone deacetylase inhibitor trichostatin A (TSA)¹⁰. To determine

whether *Apaf-1* is repressed by methylation, melanoma cells were treated with 3 μ M 5-aza2dC for 2–3 population doublings, and Apaf-1 expression was examined. 5-aza2dC treatment produced an 8- to 20-fold increase in Apaf-1 protein and RNA levels in Apaf-1-negative melanoma cells, but had little effect on Apaf-1-positive cells (Fig. 2d; Supplementary Information; data not shown). TSA also increased Apaf-1 levels in Apaf-1-negative cells, but nocodazole, an anti-proliferative drug that does not affect methylation, did not (see Supplementary Information). This indicates that transcriptional silencing by methylation may contribute to Apaf-1 inactivation in melanoma.

Epigenetic inactivation of gene expression often involves complete transcription silencing by methylation of CpG islands in promoter regions¹⁰. To characterize the Apaf-1 promoter, we sequenced roughly 5 kilobases (kb) upstream from the transcriptional starting site (see Supplementary Information). Although a CpG island was found in the Apaf-1 5' untranslated region (residues –680 to +420 of the transcription starting site), methylation-specific PCR and bisulphite sequencing excluded extensive methylation in this region (see Supplementary Information). This observation is consistent with the incomplete shutdown of Apaf-1 expression in some cell lines, and suggests that a transactivating regulator of Apaf-1 may be modulated by methylation, or, more likely, that the methylation affects an enhancer or other regulatory element outside the Apaf-1 core promoter. Such enhancer-related methylation events have been described for the imprinting of *H19* and *Igf2* genes^{11,12}. Of note, we found that the *Apaf-1* gene (> 100 kb) has at least six additional CpG islands and a high content of

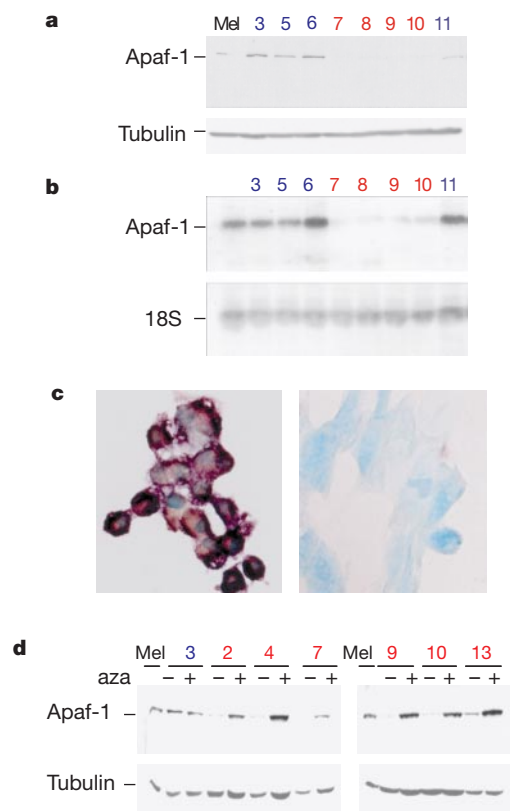


Figure 2 Apaf-1 loss in melanoma cell lines. **a**, **b**, Western blot (**a**) and northern blot (**b**) analyses of Apaf-1 expression of normal human epidermal melanocytes (Mel) and indicated melanoma cell lines. Apaf-1-positive cell lines (blue), and Apaf-1-negative cell lines (red) are indicated (see Supplementary Information). **c**, Visualization of Apaf-1 mRNA in cell lines 3 (left) and 12 (right) by *in situ* hybridization. The probe used is the same as in Fig. 1c. **d**, Effect of 5-aza2dC treatment on Apaf-1 levels in representative cell lines (see Supplementary Information).

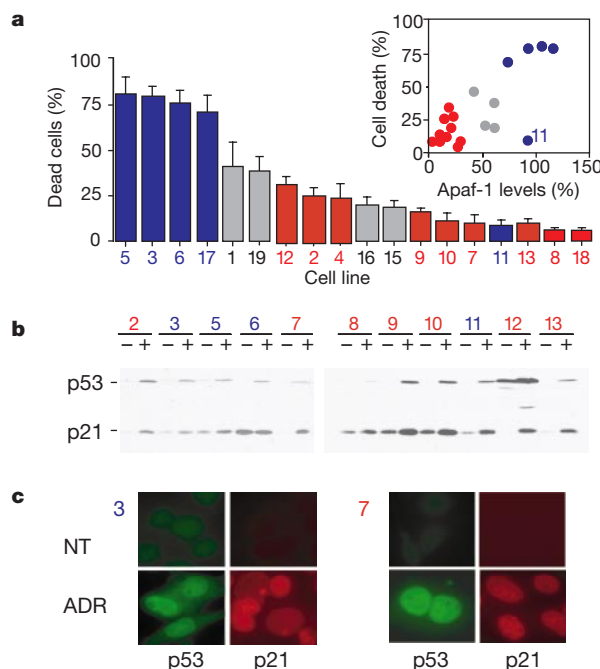


Figure 3 Apaf-1 deficiency compromises p53-dependent apoptotic response to chemotherapeutic drugs. Apaf-1-positive cell lines are indicated in blue, and Apaf-1-negative cell lines are indicated in red. Cell lines with intermediate Apaf-1 levels are indicated in black (see Supplementary Information). **a**, Chemosensitivity of melanoma cells assessed 36 h after treatment with 0.6 μ g ml⁻¹ ADR. Data are plotted as the percentage of dying cells or as relative viability as a function of Apaf-1 protein levels (inset; see Supplementary Information). Each bar represents the mean $\pm$ s.d. of at least three experiments. **b**, Induction of p53 and the p53 target, p21, on ADR treatment shown by western blots of untreated cells (–) and cells treated for 4 h with 0.6 μ g ml⁻¹ ADR. **c**, Indirect immunofluorescence of p53 and p21 from an Apaf-1-positive (3) and Apaf-1-negative (7) cell line 24 h after ADR treatment.

repetitive elements that could contribute to epigenetic gene repression (R. Davuluri, M.S., M. Zhang and S.W.L., unpublished observations).

To determine the impact of Apaf-1 loss on the p53 pathway, melanoma cells were treated with adriamycin (ADR), a chemotherapeutic agent that activates p53 to induce apoptosis⁵. At lower doses ($\leq 0.3 \mu\text{g ml}^{-1}$), all melanoma cell lines were insensitive to ADR (data not shown). At higher doses, a good correlation was observed between Apaf-1 levels and ADR-induced death, with all Apaf-1-negative cells being extremely resistant (Fig. 3a). Only one cell line (11) was Apaf-1-positive and highly resistant to ADR. Notably, the few Apaf-1-negative cells that died showed incomplete chromatin condensation, which was in contrast to the typical apoptotic chromatin condensation observed in Apaf-1-positive cells (see below). Both Apaf-1-positive and -negative cells retained a similar capacity to activate p53, as ADR induced p53 and its transcriptional target p21 in the nucleus of most cell lines analysed (Fig. 3b, c; and Supplementary Information). Ectopic p53 expression by adenovirus-mediated gene transfer induced cell death in Apaf-1-positive cells but had little effect in Apaf-1-negative cells (Fig. 4a and b). This suggests that Apaf-1 negative melanoma cells have an effector defect in the p53 apoptotic programme.

The adenovirus *E1A* oncogene induces apoptosis through p53 and Apaf-1 (refs 6, 7 and 13), and BAX (a p53 effector) facilitates cytochrome *c* release, which leads to activation of the Apaf-1/Casp9 death-effector complex¹⁴. To determine the impact of these genes on melanoma cell viability, *E1A* and BAX were introduced into several Apaf-1-positive and -negative cell lines (Fig. 4c). Melanoma cells had similar responses to *E1A* and to ADR. Similarly, Apaf-1-negative cell lines were resistant to BAX, whereas all Apaf-1-positive cell lines, including cell line 11, were sensitive. The difference in sensitivity of line 11 to *E1A* and BAX is consistent with a defect

between p53 and the mitochondria¹⁴. However, all Apaf-1 negative melanoma cells were resistant to stimuli that induce cell death through Apaf-1.

Cytochrome *c* is normally sequestered in the mitochondria, but it is released into the cytosol on treatment with certain apoptotic stimuli. There, cytochrome *c* triggers oligomerization of Apaf-1, which then binds and activates Casp9 (refs 15–17). These events can be monitored by cell fractionation or immunolocalization of cytochrome *c*, or by immunoblotting to detect the processed form of Casp9. As expected, cell line 11 retained cytochrome *c* on treatment with ADR (data not shown). In other cell lines examined, ADR induced a slow release of cytochrome *c* that was independent of Apaf-1 status (Fig. 5a and b). Nevertheless, ADR was very inefficient in promoting Casp9 processing in Apaf-1-negative cells (Fig. 5c). Notably, no Casp9 mutations were detected in eight cell lines examined (3, 5–11) (data not shown). Therefore, Apaf-1-negative melanoma cells contain a biochemical lesion that uncouples cytochrome *c* release from Casp9 processing.

To show that Apaf-1 is functionally inactivated in melanoma, we re-expressed Apaf-1 using retroviral-mediated gene transfer or by treatment with 5-aza2dC. Ectopic expression of Apaf-1 in negative cell lines restored Apaf-1 expression to physiological levels (Fig. 6a), and rescued the defect in Casp9 processing and chromatin condensation (Fig. 6b and c). Moreover, these cells regained a level of chemosensitivity comparable to Apaf-1-positive cells (Fig. 6d). Similarly, recovery of Apaf-1 expression by 5-aza2dC treatment enhanced the chemosensitivity of Apaf-1-negative cells (Fig. 6e; see Fig. 3a for comparison). Neither ectopic expression of *Apaf-1* nor 5-aza2dC treatment induced a significant increase of ADR-induced apoptosis in Apaf-1-positive cells.

These results show that Apaf-1 is inactivated in some melanomas and suggest that Apaf-1 is a tumour suppressor. In melanoma,

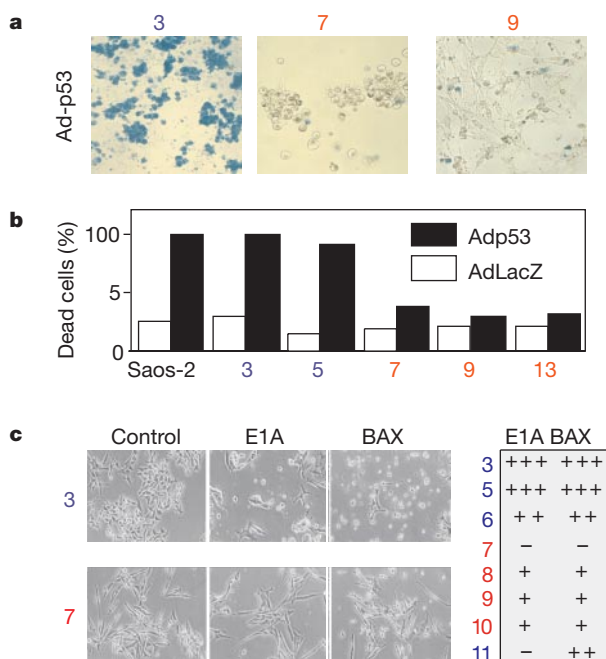


Figure 4 Apoptotic defects in Apaf-1-negative melanoma cells. **a**, Viability measurements of melanoma cells after infection with a replication-defective adenovirus expressing p53 (Adp53) or a control LacZ. Seventy-two hours after infection, representative cultures were stained with trypan blue to visualize dead cells. **b**, Quantification of data in **a**. Saos-2 cells are included as a positive control²⁸. **c**, Viability of representative melanoma lines 36 h after introduction of *E1A* or BAX by retrovirus-mediated gene transfer. Photomicrographs of two representative cell lines are indicated, with quantification of data from a larger series shown to the right. The colour code for the cell lines is the same as in Fig. 3.

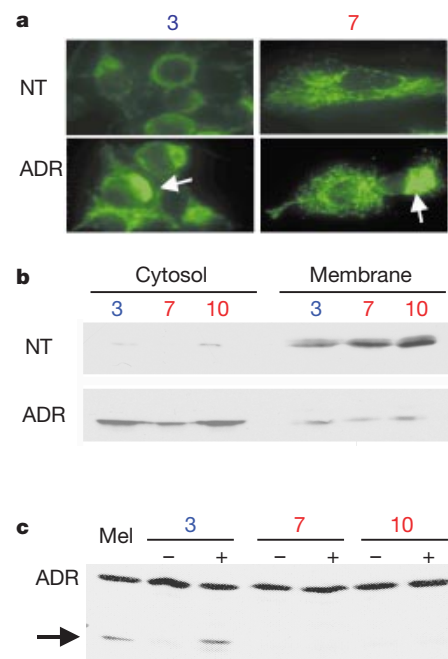


Figure 5 Apaf-1 loss uncouples cytochrome *c* release from Casp9 activation in melanoma cells. **a**, Immunolocalization (confocal microscopy) of cytochrome *c* in non-treated (NT) or ADR-treated cells for 24 h. White arrows indicate cells with cytoplasmic cytochrome *c*. **b**, Protein immunoblot of 50 μg soluble, cytosolic proteins and 10 μg membrane-bound proteins, showing the differential localization of cytochrome *c* before and 48 h after ADR treatment. **c**, Casp9 activation assessed by immunoblotting in untreated (–) and ADR-treated (36 h) normal melanocytes (mel) and the indicated melanoma cell lines. Arrow indicates the processed form of Casp9.

inactivation of *Apaf-1* apparently arises from LOH and from gene silencing by methylation. Like other methylated genes^{8,9}, no point mutations were detected in the *Apaf-1* coding sequence. Thus, *Apaf-1* joins a growing list of cancer genes that can be inactivated by epigenetic mechanisms^{8,9,18}. It is noteworthy that heterozygous *Apaf-1* frameshift mutations have been reported in a small percentage of mismatch repair-defective gastrointestinal cancers, although the functional significance of these alterations is unknown¹⁹. In

melanoma, *Apaf-1* is clearly a functional target of lesions on chromosome 12, as LOH is accompanied by loss of *Apaf-1* expression, and re-expression of *Apaf-1* reverts the apoptotic defects associated with *Apaf-1* loss. Therefore, although our data do not exclude the possibility that additional genes are targeted by losses on chromosome 12q22–23, they establish the importance of the *Apaf-1*/Casp9 death-effector complex in tumour control and imply that this complex can be rate limiting in cell death.

These results have marked implications for understanding the origins and behaviour of metastatic melanoma. They show that a substantial percentage of melanomas inactivate *Apaf-1* and, as a consequence, disable the p53 apoptotic programme. Although *Apaf-1* can process p53-independent signals, the phenotypic similarities between melanomas and tumours with high p53 mutation rates suggest that p53 dysfunction is a biologically relevant outcome of *Apaf-1* loss. *INK4a/ARF* deletions (see Supplementary Information), which can also disable p53 (refs 4 and 20) and contribute to melanoma^{1,21,22}, may further decrease the selective pressure to mutate p53, such that the p53 pathway may be lost in most melanomas. Our results also imply that *Apaf-1* loss contributes to the aggressive nature and extreme chemoresistance of metastatic melanoma. Although other anti-apoptotic lesions may also be involved^{23–25}, we have shown that *Apaf-1* is required for proper apoptosis in melanoma cells, and that restoring *Apaf-1* to physiological levels enhances chemosensitivity. Assessment of *Apaf-1* status may therefore improve the therapeutic management of patients with malignant melanoma. □

Methods

Analysis of the *Apaf-1* locus

For LOH analysis, frozen-tumour specimens with greater than 70% tumour cellularity were selected for analysis. Genomic DNA (50–100 ng) from matched tumour and normal DNA was amplified by PCR and analysed as described²⁶ using fluorescently labelled primers for the indicated polymorphic microsatellite markers on chromosome 12q22–23 (NIH Genemap99).

Cells and gene transfer

Melanoma lines analysed are described in Supplementary Information. All cell lines were maintained in DMEM supplemented with 10% FBS. We cultured normal melanocytes as suggested by the provider (Clonetics). We expressed the murine ecotropic retrovirus receptor in each melanoma line using a retroviral vector (pWZL/Neo/Eco) essentially as described²⁷. Infected cells were isolated by selection in 500–800 $\mu\text{g ml}^{-1}$ G418. All subsequent infections with ecotropic viruses were performed as described^{3,27} using an empty vector control (pBabe/puro), or vectors expressing BAX (pBabe/puro/BAX), E1A (pBabe/puro/13S) or *Apaf-1* (pBabe/puro/*Apaf-1*). Adenovirus vectors were used at a multiplicity of infection of 100 (ref. 28). For methylation studies, cells were incubated with 3 μM 5aza2dC (Sigma), while maintaining subconfluent density, for 3–5 days.

Gene expression

Whole-cell extracts²⁰ (15–30 μg) were resolved on polyacrylamide–SDS gels, and transferred to Immobilon-P membrane (Amersham). Proteins were detected using enhanced chemiluminescence (ECL), using antibodies directed against the amino terminus of *Apaf-1* (monoclonal 6-1-19 and polyclonal 8.46, 1:1,000 dilution), p53 (CM1, Novocastra, 1:1,000 dilution), p21 (C-19, Santa Cruz, 1:1,000 dilution), Casp9 (96-2-22, 1:1,000 dilution) or tubulin (B-5-1-2, Sigma, 1:2,000 dilution). For northern blots, 20 μg of total RNA was separated on a 1% agarose–formaldehyde gel, transferred to Hybond-N membrane (Amersham), and probed with radioactively labelled complementary DNA fragments corresponding to nucleotides 272–2,908 and 1,517–3,525. We performed *in situ* hybridization using sense and antisense riboprobes containing *Apaf-1* fragments corresponding to nucleotides 1–461, 581–1,363, 2,112–2,225 or 3,426–3,710, labelled with digoxigenin-UTP. Cell cytopins or deparaffinized tissue sections were mounted on coverslips and processed as described²⁹. Hybridizations contained 10 pmol l^{-1} labelled probe in 50 μl of hybridization buffer (50% de-ionized formamide (v/v), 10% dextran sulphate, 2 $\times$ SSC, 1% SDS and 0.25 mg ml^{-1} of herring sperm DNA), and were performed at 45 $^{\circ}\text{C}$ overnight.

Analysis of cell death

We determined cell viability, chromatin condensation and Casp9 processing by trypan blue exclusion, DAPI fluorescence and immunoblotting, respectively^{3,15,27}. Cell viability in the presence of 5aza2dC was determined: 2×10^6 cells (two plates each) were placed in 1 μM 5aza2dC, and 48 h later medium was replaced with fresh 5aza2dC (1 μM) in the absence or presence of 0.3 μM ADR. We determined viability after 24 h. Release of cytochrome *c* from mitochondria was measured on treatment with 0.6 $\mu\text{g ml}^{-1}$ ADR by

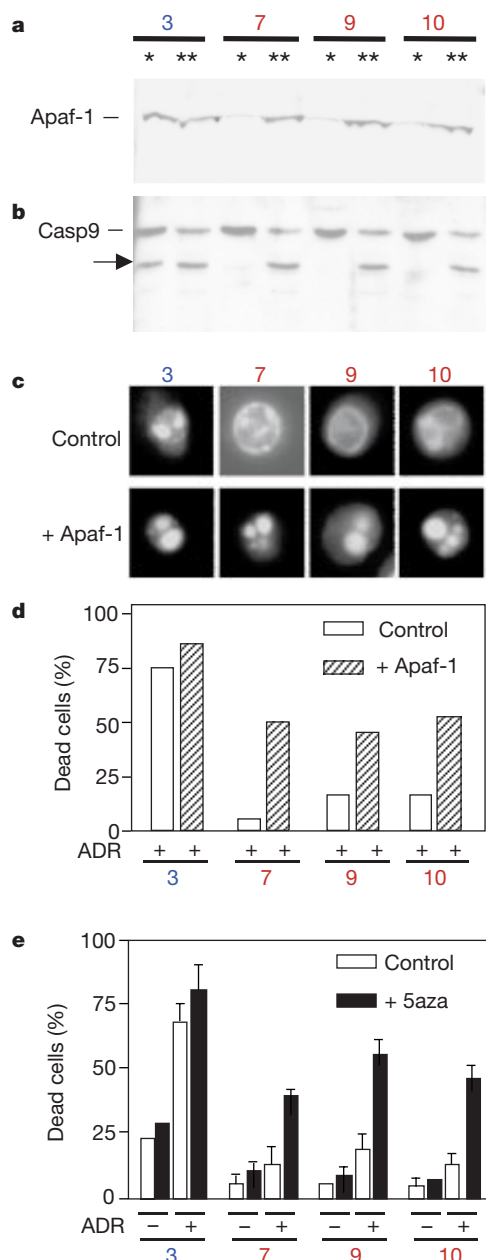


Figure 6 Rescue of apoptotic defects by re-expression of *Apaf-1*. **a**, *Apaf-1* levels in melanoma cells infected with control vector (asterisk) and *Apaf-1*_{XL}-expressing retrovirus (double asterisk). **b**, Cells were treated with 0.4 $\mu\text{g ml}^{-1}$ ADR and collected 36 h later to analyse Casp9 processing (arrow) by immunoblotting, and to visualize chromatin condensation. **c**, DAPI staining of floating (dead) treated cells showing different chromatin condensation pattern of control cells (top) or cells expressing ectopic *Apaf-1* (bottom). The amount of dead cells with fully condensed chromatin increased from about 30% to 70% on *Apaf-1* expression. **d**, Viability of control cells or *Apaf-1*-retroviruses treated as in **c**. **e**, Viability of cells untreated (control) or treated with 1 μM 5aza2dC in the absence (–) or in the presence (+) of 0.3 μM ADR.

indirect immunofluorescence using an anti-cytochrome *c* antibody (65972A, PharMin-gen, 1:1,000), or by cell fractionation into cytosol and membrane components and immunoblotting (cytochrome *c* antibody 7H8.2C12, PharMingen, 1:1,000 dilution).

Received 29 June; accepted 10 November 2000.

- Chin, L., Merlino, G. & DePinho, R. A. Malignant melanoma: modern black plague and genetic black box. *Genes Dev.* **12**, 3467–3481 (1998).
- Serrone, L. & Hersey, P. The chemoresistance of human malignant melanoma: an update. *Melanoma Res.* **9**, 51–58 (1999).
- Soengas, M. S. *et al.* Apaf-1 and caspase-9 in p53-dependent apoptosis and tumor inhibition. *Science* **284**, 156–159 (1999).
- Schmitt, C. A. & Lowe, S. W. Apoptosis and therapy. *J. Pathol.* **187**, 127–137 (1999).
- Lowe, S. W., Ruley, H. E., Jacks, T. & Housman, D. E. p53-dependent apoptosis modulates the cytotoxicity of anticancer agents. *Cell* **74**, 957–967 (1993).
- Hakem, R. *et al.* Differential requirement for caspase 9 in apoptotic pathways *in vivo*. *Cell* **94**, 339–352 (1998).
- Fearnhead, H. O. *et al.* Oncogene-dependent apoptosis is mediated by caspase-9. *Proc. Natl Acad. Sci. USA* **95**, 13664–13669 (1998).
- Baylin, S. B. & Herman, J. G. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet.* **16**, 168–174 (2000).
- Robertson, K. D. & Jones, P. A. DNA methylation: past, present and future directions. *Carcinogenesis* **21**, 461–467 (2000).
- Bird, A. P. & Wolffe, A. P. Methylation-induced repression—belts, braces, and chromatin. *Cell* **99**, 451–454 (1999).
- Hark, A. T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486–489 (2000).
- Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482–485 (2000).
- Lowe, S. W. & Ruley, H. E. Stabilization of the p53 tumor suppressor is induced by adenovirus 5 E1A and accompanies apoptosis. *Genes Dev.* **7**, 535–545 (1993).
- Green, D. R. & Reed, J. C. Mitochondria and apoptosis. *Science* **281**, 1309–1312 (1998).
- Rodriguez, J. & Lazebnik, Y. Caspase-9 and APAF-1 form an active holoenzyme. *Genes Dev.* **13**, 3179–3184 (1999).
- Zou, H., Li, Y., Liu, X. & Wang, X. An APAF-1 cytochrome *c* multimeric complex is a functional apoptosome that activates procaspase-9. *J. Biol. Chem.* **274**, 11549–11556 (1999).
- Hu, Y., Benedict, M. A., Ding, L. & Nunez, G. Role of cytochrome *c* and dATP/ATP hydrolysis in Apaf-1-mediated caspase-9 activation and apoptosis. *EMBO J.* **18**, 3586–3595 (1999).
- Teitz, T. *et al.* Caspase 8 is deleted or silenced preferentially in childhood neuroblastomas with amplification of MYCN. *Nature Med.* **6**, 529–535 (2000).
- Yamamoto, H., Gil, J., Schwartz, S. Jr & Peruchio, M. Frameshift mutations in Fas, Apaf-1, and Bcl-10 in gastro-intestinal cancer of the microsatellite mutator phenotype. *Cell Death Differ.* **7**, 238–239 (2000).
- Schmitt, C. A., McCurrach, M. E., de Stanchina, E., Wallace-Brodeur, R. R. & Lowe, S. W. INK4a/ARF mutations accelerate lymphomagenesis and promote chemoresistance by disabling p53. *Genes Dev.* **13**, 2670–2677 (1999).
- Chin, L. *et al.* Cooperative effects of INK4a and ras in melanoma susceptibility *in vivo*. *Genes Dev.* **11**, 2822–2834 (1997).
- Walker, G. J. *et al.* Virtually 100% of melanoma cell lines harbor alterations at the DNA level within CDKN2A, CDKN2B, or one of their downstream targets. *Genes Chromosom. Cancer* **22**, 157–163 (1998).
- Grossman, D., McNiff, J. M., Li, F. & Altieri, D. C. Expression and targeting of the apoptosis inhibitor, survivin, in human melanoma. *J. Invest. Dermatol.* **113**, 1076–1081 (1999).
- Irmir, M. *et al.* Inhibition of death receptor signals by cellular FLIP. *Nature* **388**, 190–195 (1997).
- Selzer, E. *et al.* Expression of Bcl-2 family members in human melanocytes, in melanoma metastases and in melanoma cell lines. *Melanoma Res.* **8**, 197–203 (1998).
- Mora, J. *et al.* Clinical categories of neuroblastoma are associated with different patterns of loss of heterozygosity on chromosome Arm 1p. *J. Mol. Diag.* **2**, 1–9 (2000).
- Woo, M. *et al.* Essential contribution of caspase 3/CPP32 to apoptosis and its associated nuclear changes. *Genes Dev.* **12**, 806–819 (1998).
- Marcellus, R. C., Teodoro, J. G., Charbonneau, R., Shore, G. C. & Branton, P. E. Expression of p53 in Saos-2 osteosarcoma cells induces apoptosis which can be inhibited by Bcl-2 or the adenovirus E1B-55 kDa protein. *Cell Growth Differ.* **7**, 1643–1650 (1996).
- Capodice, P., Magi-Galluzzi, C., Moreira, G. Jr Zeheb, R. & Loda, M. Automated *in situ* hybridization: diagnostic and research applications. *Diagn. Mol. Pathol.* **7**, 69–75 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank B. Stillman and N. K. Cheung for support; M. McCurrach, W. Jiang, L. Chen and M. Dudas for technical assistance; M. Zhang, R. Davuluri and I. Ioshikhes for analysis of Apaf-1 CpG islands and assembly of Apaf-1 genomic sequence; G. Núñez for Apaf-1_{XL} pcDNA construct; E. Querido, P. E. Branton and F. L. Graham for Ad-p53 and Ad-LacZ; and A. Houghton and R. Camalier for melanoma cell lines. We also thank G. Hannon, C. Schmitt, C. Hazan, K. Pohar, T. Tiongsong and the Cold Spring Harbor Genome Center for help and advice. Y.L. is a Pew Scholar and S.W.L. is a Rita Allen Scholar. This work was supported by a long-term Postdoctoral Fellowship from the Human Frontiers in Science Program and a Special Fellowship from the Leukemia and Lymphoma Society (M.S.S.), a K08 grant (D.P.), an ASCO Young Investigator Award (J.M.), grants from the Seraph Foundation (Y.L.), and from the NIH and the NCI (C.C.C. and S.W.L.).

Correspondence and requests for materials should be addressed to S.W.L. (e-mail: lowe@cshl.org).

The protein–protein interaction map of *Helicobacter pylori*

Jean-Christophe Rain*, Luc Selig*, Hilde De Reuse†, Véronique Battaglia*, Céline Reverdy*, Stéphane Simon*, Gerlinde Lenzen*, Fabien Petel*, Jérôme Wojcik*, Vincent Schächter*, Y. Chemama*, Agnès Labigne† & Pierre Legrain*

* Hybrigenics SA, 180 avenue Daumesnil, Paris 75012, France

† Unité de Pathogénie Bactérienne des Muqueuses, Institut Pasteur, 25 rue du Dr Roux, 75724 Paris Cedex 15, France

With the availability of complete DNA sequences for many prokaryotic and eukaryotic genomes, and soon for the human genome itself, it is important to develop reliable proteome-wide approaches for a better understanding of protein function¹. As elementary constituents of cellular protein complexes and pathways, protein–protein interactions are key determinants of protein function. Here we have built a large-scale protein–protein interaction map of the human gastric pathogen *Helicobacter pylori*. We have used a high-throughput strategy of the yeast two-hybrid assay to screen 261 *H. pylori* proteins against a highly complex library of genome-encoded polypeptides². Over 1,200 interactions were identified between *H. pylori* proteins, connecting 46.6% of the proteome. The determination of a reliability score for every single protein–protein interaction and the identification of the actual interacting domains permitted the assignment of unannotated proteins to biological pathways.

During the past few years, interaction maps have been proposed for viral^{3–5} and eukaryotic (*Saccharomyces cerevisiae*^{2,6,7} and *Caenorhabditis elegans*⁸, for example) genomes. Here we describe the first procaryotic interaction map, for which the strategy used (Fig. 1) is also a variation of the two-hybrid assay². It differs considerably by both the type of results generated and the throughput of the experimental procedures. We constructed a library of random genomic fragments of the *H. pylori* strain 26695 that had been previously sequenced⁹. A high complexity library was first obtained in *Escherichia coli* (over ten million clones). Ninety-seven per cent of the plasmids contained a single genomic insert (mean size 1,000 ± 550 nucleotides). This library was then introduced into yeast by transformation. Two million independent yeast colonies were collected, pooled and stored at –80 °C as equivalent aliquot fractions of the same library.

In parallel, a large set of bait plasmids was constructed in a bait vector designed to decrease the level of transcriptional auto-activation². Bait constructs were specifically adapted for interaction screens, yielding 'validated baits'; for example, hydrophobic putative trans-membrane domains were discarded to avoid any non-nuclear localization of the bait protein in the yeast cell. In some cases, specific open reading frame (ORF) domains were selected for the bait design. For every single bait construct, we performed a preliminary small-scale screening experiment that determines the selective pressure (that is, modifying the selective medium, see Methods) to be applied to obtain a small number of independent positive clones per million interactions tested (usually less than 10). All positive colonies were then picked, and prey fragments were individually identified by sequence analysis and comparison with the genomic database through a dedicated integrated laboratory production management system (Fig. 1).

Protein–protein interaction maps are built on experimental data that ideally yield a heuristic value for each connection. Our procedure involves several steps of processing of raw two-hybrid results (Fig. 1). First, positive prey fragments are clustered into families of overlapping fragments. The common sequence shared by

these fragments is referred to as the selected interacting domain (SID). Second, SIDs that do not code for part of an *H. pylori* ORF are discarded. Third, for every remaining SID, a PIM biological score (PBS) is computed. The PBS is based on a statistical model of the competition for bait-binding between fragments. It is computed like a classical expected value (*E* value), and ranges from 0 (specific interaction) to 1 (probable artefact). For practical use, the scores were divided into four categories, from A (score very close to 0) to D (close to 1). A fifth category, E, was added to distinguish interactions involving only highly connected prey domains (SIDs found as prey with frequency greater than a fixed threshold). These are most probably two-hybrid artefacts. Although they may have some biological significance, they add little specific information to the interaction map. It should be emphasized that domains, rather than whole proteins, are tagged. For example, the carboxy-terminal region of protein HP0705/UvrA was found as a prey in 16 different interactions that were scored in the E category, but the same protein was selected through another domain as a specific interacting prey. Because global connectivity is taken into account, the PBS is

computed incrementally over the whole PIM and its discriminatory power increases as screening results accumulate.

We carried out 285 screens on 261 *H. pylori* bait ORFs chosen as follows: first, a core set of 50 proteins known to be involved in complexes and/or in pathogenicity was used to validate the approach; and second, 211 baits were then picked randomly, with a slight bias toward regions of the proteome that were still unexplored. Positive colonies (13,962) were selected (Table 1), and more than 95% of them were identified by sequencing the prey insert. From these prey fragments, 2,680 independent SIDs were defined, out of which 1,100 fell into non *H. pylori* ORF coding regions and were discarded (Fig. 1). Thirty-one SIDs were classified in the E category. The remaining SIDs define 1,280 interactions including 62 homo-oligomeric interactions. In total, 46.6% *H. pylori* proteins of the proteome were connected, which corresponds to an average connectivity of 3.36 partners per connected protein (without counting homodimeric connections). Only 14 screens out of 285 yielded no positive clones or interactions with nonsignificant score values, illustrating the fact that our technology reduces the rate of false negatives.

Given that little information is available on protein interactions in *H. pylori*, we used a data set of known interactions in *E. coli* to validate further our experimental strategy, and to evaluate the correlation between the PBS and the actual biological significance of interactions. For each *H. pylori* protein present in the interaction map, significant *E. coli* orthologues (FastA score < 0.01) were selected and their annotations in the SwissProt database (release 38)¹⁰ were verified manually for known interactions shown by various biochemical means. The resulting *E. coli* interaction list was compared with interactions found in *H. pylori* (Fig. 2). Among these *E. coli* interacting pairs, 53% of the homodimers and 67% of the heterodimers were found. Among heterodimers, five out of six negative pairs were tested only in one direction, suggesting that performing the reciprocal screens would decrease this number. Most interactions that were described for orthologous proteins in *E. coli* fell into the high-scoring interaction category (A) according to our PBS calculation (7/10 homodimers and 9/12 heterodimers) confirming the heuristic value of the classification. The interaction map was also analysed according to the classification of *H. pylori* proteins into 14 functional categories previously proposed⁹. For 10 categories out of 14, more intra-category protein–protein interactions were observed than expected from a random theoretical distribution (*Z*-scores ranging from 2 to 50; in all cases, *P* < 0.05), suggesting the existence of a significant correlation between functional grouping and detection of interactions.

To display and analyse the interaction data, we developed a software platform composed of a database, a web-based graphical interface layer and various query and analysis tools (the PIMRider, Fig. 3; see also <http://pim.hybrigenics.com>). Starting from a gene name (or an ORF name), the PIMRider draws an automatic layout of the neighbourhood of this protein in the protein interaction map (Fig. 3a). Paths connecting two proteins can also be queried.

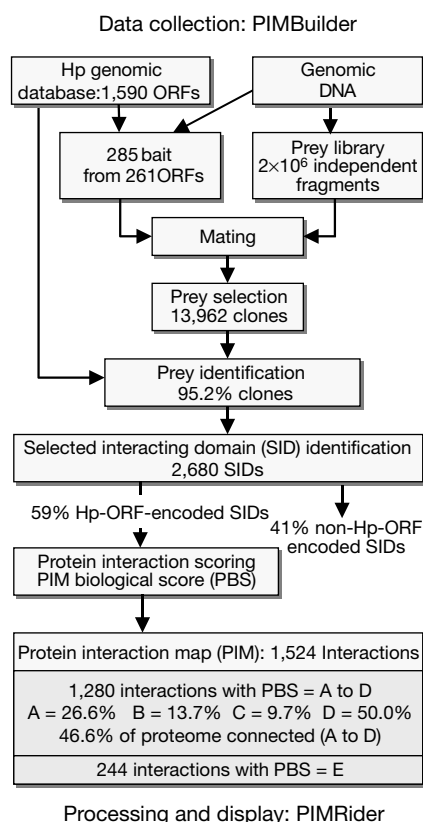


Figure 1 Outline of the strategy for building an *H. pylori* (Hp) proteome-wide interaction map. A production database (the PIMBuilder) was built that contains information related to the genomic sequence of *H. pylori*, which codes for 1,590 putative proteins or ORFs. It was populated with raw data from screening experiments. The PIMBuilder tracks all biotechnological or bioinformatics operations performed during the production processes, stores information about all biological objects produced during experiments, and interfaces with robots and bioinformatics modules. It also implements the procedure used to construct PIMs from raw experimental data. After identification of almost all positive clones, overlapping prey fragments were clustered into families to define SIDs. Those families that had no biological coding capability (antisense or intergenic region, out of frame fragments occurring in a single frame) were discarded. A PIM biological score (PBS; see Methods) was then calculated for *H. pylori* ORF-encoded SIDs. Interactions were grouped into categories A to D (from high to low heuristic values). The global connectivity of the PIM was also analysed to detect highly connected prey polypeptides. Those interactions were grouped in the E category. Processing of data and visualization of interactions were performed by an in-house bioinformatic platform (PIMRider).

Table 1 General features of the *H. pylori* interaction screens

Screening data	Total or mean value	Comments
Number of screens	285	261 different <i>H. pylori</i> ORFs
Combination of bait/prey polypeptides assayed	5.6 × 10 ⁹	
Number of prey polypeptides assayed per bait	2 × 10 ⁷	
Number of selected colonies per bait (1st reporter)	916	From 0 to 15,000
Number of selected clones per bait (2 reporters)	49	From 0 to 139
Number of selected colonies per million pairs tested	3.8	From 0 to 25
Number of selected clones	13,962	
Number of sequenced clones	13,296	95.2 %

Connections are displayed with their PBS scores, and can be filtered according to score categories. A graphical summary of information describing all interaction domains within a given protein can be displayed (Fig. 3b). Raw data on every interaction can be retrieved. Finally, the PIMRider supplies a description of each gene with functional and genomic information, and includes links to significant bibliographic references and to relevant external databases, such as PyloriGene (<http://genolist.pasteur.fr/PyloriGene/>). PyloriGene is a manually annotated database of the two *H. pylori* sequenced genomes^{9,11}, that integrates all publicly available information on genes and proteins and has been elaborated with a structure similar to that of the *Bacillus subtilis* SubtiList database¹².

Exploring the protein–protein interaction map reveals biological pathways and allows prediction of protein function. A first example concerns chemotaxis (see Fig. 3a). The *H. pylori* genome reveals three homologues of *E. coli* proteins that are involved in the chemotactic pathway (CheA, CheW and CheY) and proteins such as TlpA similar to chemotaxis receptors (MCPs). CheA was found to interact with CheY and CheW, and distinct interacting domains were precisely identified (Fig. 3b). The domain of CheA which binds to CheY precisely overlaps with the interacting domain assigned by a structural study in *E. coli*¹³. The TlpA-binding site for CheW was localized in a domain known, in *E. coli*, to be methylated and implicated in the transduction of the chemotactic signal.

The urease complex was also examined. Urease activity is essential for *H. pylori* pathogenicity and its synthesis requires two structural subunits, UreA and UreB, and the product of four accessory genes: *ureE*, *ureF*, *ureG*, *ureH*¹⁴. Complexes between accessory proteins and their role in nickel incorporation at the urease active site have been described for orthologues, but little information is available for *H. pylori* (for review, see ref. 15). The

protein interaction map revealed the connection between UreA and UreB, and one of the two expected homo-oligomeric interactions of structural subunits (UreA); the UreB homodimer could not be detected. A connection between accessory proteins and the structural subunits occurs via UreH and UreA, which is consistent with the presumed chaperone role of UreH¹⁵. A new structural link was found between UreG and UreE. The UreF and UreH proteins were connected, but no connection between UreG and UreF or UreH was detected. In addition to the accessory proteins, the urease operon codes for an inner-membrane protein, UreI, essential for resistance to acidity¹⁶ and recently described as a H⁺-gated urea channel¹⁷. The third cytoplasmic domain of this protein reveals a potential interaction with the ExbD protein which is involved in transmission of PMF (proton motor force) energy to outer-membrane receptors.

Combination of genomic and proteomic data also permits function prediction. The *H. pylori* proteome contains a homologue of the *E. coli* HolB protein. In *E. coli*, this protein interacts with HolA to form part of the DNA polymerase core¹⁸. We found one high-scoring interaction between *H. pylori* HolB and an uncharacterized polypeptide, HP1247. A pairwise alignment between *E. coli* HolA and HP1247 highlighted structural homology (Fig. 4) not found by previous sequence analysis (Fig. 2). We thus assign the HolA function to the *H. pylori* HP1247 protein. The organization of bacterial genomes into operons suggests a functional relationship between the corresponding gene products that can be directly compared with our protein interaction map. Indeed, we found interactions between proteins that were likely to be expressed from a single operon (Table 2). Among these, we detected interactions between proteins known to interact in *H. pylori* (ScoA–ScoB) or in other organisms (RpsR–RpsF, MoaE–MoaD, FtsA–FtsZ), and between polypeptides involved in the same enzymatic activity and not yet described as interacting (HypE–HypF).

Selected interacting domains can also be analysed in terms of protein structure. The prokaryotic RNA polymerase, composed of a core enzyme ($\alpha_2\beta\beta'$) associated with a σ -factor, is one of the best studied multisubunit enzymes. In *H. pylori*, the β - and β' -subunits

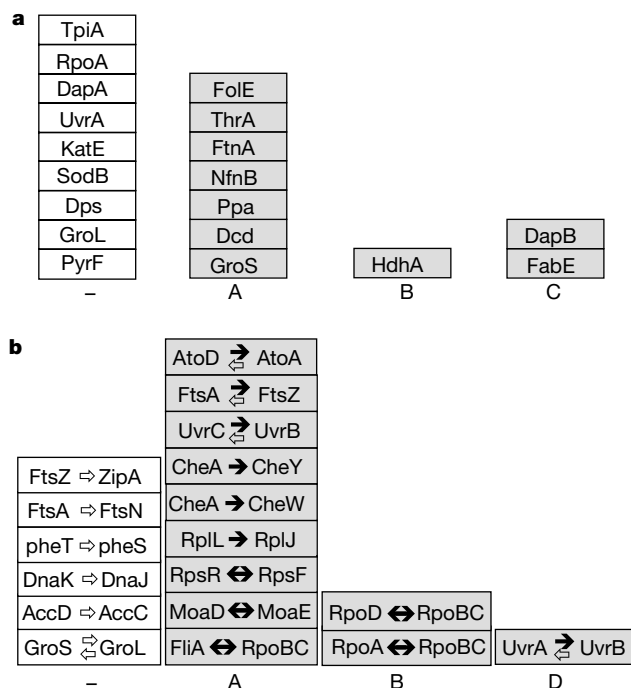


Figure 2 Sets of *E. coli* interaction data for which *H. pylori* orthologous proteins were identified and assayed in interaction screens. **a**, Homodimers; **b**, heterodimers. Names of *E. coli* proteins are boxed. Unidentified interactions between *H. pylori* orthologues are scored as ‘–’ and shown in white boxes. Identified *H. pylori* interactions are indicated in grey boxes with their PBS score category (A, B, C or D). When several orthologues were identified, only the best scoring homologue was considered. For heterodimers, arrows indicate the direction of the screen (bait to prey) that was performed. The black (or white) colour of the arrow indicates a positive (or negative) result in the interaction screen.

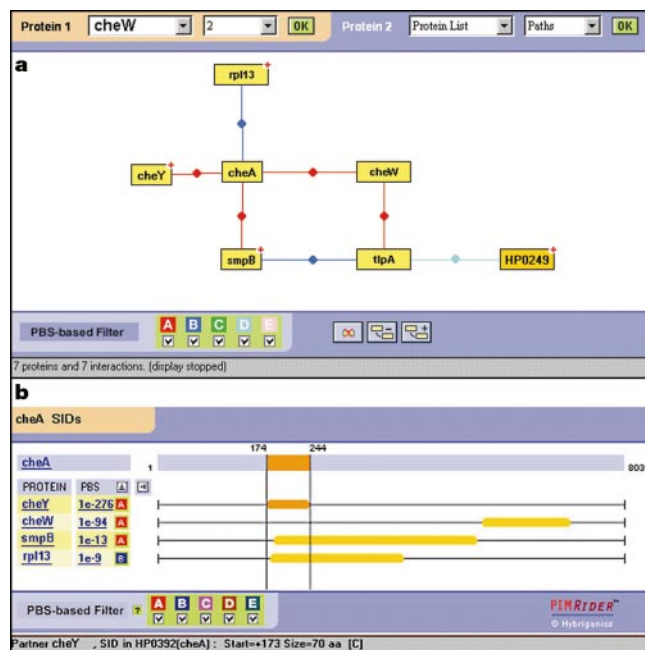


Figure 3 PIMRider screen shots. **a**, The PIMViewer displays a portion of the protein interaction map around the CheA protein. Links between proteins identify connections with their colour-coded PBS score; **b**, the MultiSID Viewer exhibits the various interacting domains in the CheA protein (for details see <http://pim.hybrigenics.com>).

usually found in other bacteria are fused into a single polypeptide (RpoB). One of the two alternative σ -factors present in *H. pylori*, (HP1032), is similar to the *E. coli* FliA protein (sigma 28) necessary for transcription of genes involved in flagellar biosynthesis¹⁹. We identified a precise region of RpoB that interacts with the *H. pylori* FliA. This selected interacting domain (residues 841–959) maps exactly to a structural domain called the flexible flap²⁰. The RpoB-interacting domain of FliA falls in the regions 3.2–4.2 of this σ -factor (residues 175–255). Biochemical studies suggest an interaction between the flap domain of RpoB and region 4 of sigma factors²⁰. Our experiments thus characterize this interaction and support the role of the flap domain and σ -factors in the transition from an open complex to a processive elongation complex²¹.

Our work provides a way to characterize proteome-wide protein interaction maps. Our results identify complexes that have been shown or postulated to exist in other organisms, such as in *E. coli*. They also complement sequence information about homologous proteins and operon prediction from the location of genes on the chromosome. Finally, identified interacting domains could be mapped on three-dimensional structures of proteins. As a whole they lead to the assignment of a functional role for many as yet uncharacterized proteins and provide tools, such as interacting domains, for further biological experiments. Our technology was designed to specifically address the main known causes for false-negative and false-positive results in two-hybrid assays. It is now clear that interactions are not necessarily detected positive by two-hybrid assays when used in reciprocal directions (Fig. 2; see also refs 4, 8, 22). Parallel screening against highly complex libraries of

fragments increases the number of 'two-hybrid-capable' candidates and reduces the rate of false negatives that arise with the classical two-hybrid matrix approach (that is, pair wise testing of a collection of proteins^{6,7}). Concerning the false positives, the specific design of selection procedures that permit a strong selectivity for all baits and the statistical analysis made possible by the experimental procedure (that is, reproducible exhaustive screening of fragment libraries) allows us to detect and tag nonspecific partners through a global scoring scheme.

Ultimate validation of biological significance should come from additional biological information, but comparison of our results with previously described interactions of *H. pylori* proteins and also of orthologous *E. coli* proteins supports the reliability of the approach. Protein interaction maps can be built at the scale of a proteome. Our technology is applicable to higher eukaryotes, for which highly complex random-primed complementary DNA libraries are screened for interacting domains. The identification of interacting domains is a direct consequence of the library screening approach and presents key advantages such as mapping of new functional domains or correlation between sequence similarity and functional homology. These interacting domains also constitute a first step towards the construction of dominant-negative mutants or the development of an assay for interaction modulation, applicable to new drug design. □

Methods

Bait cloning

Baits were constructed by PCR amplification and cloning in the pB6 plasmid derived from the original pAS2 $\Delta\Delta$ (ref. 2). Design of the primers was automatically proposed by a software and validated for each ORE. PCR fragments were cloned by classical enzymatic methods in a 96-well-plate format. All bait constructs used for interaction screens were fully sequenced and compared to the genome sequence; any mutant clone was discarded.

Library construction

We extracted genomic DNA from *H. pylori* 26695 as described²³, and nebulized and blunted it with a cocktail of mung bean nuclease, T4 and Klenow polymerase (NEB). Adapters containing SfiI sites were ligated to blunt DNA. The adapted DNA was cloned into pP6 plasmid derived from the original pACT2 and transformed in *E. coli* (DH10B; Life Technologies). Sequence analysis was performed on one hundred randomly chosen clones to establish the general characteristics of the library.

Screening procedure

The principle of the technique has been described². Briefly, the screening conditions were adapted for each bait during a test screen, before performing the full-size screening experiment. The selectivity of the HIS3 reporter was modulated with 3-aminotriazole (3AT). For about 15% of the screens, diploid cells were plated on selective medium containing 3AT. In 44% of the screens, the second reporter (*lacZ*) was used directly on plates for the selection of clones. In other cases, the *lacZ* reporter was used as a second round of selection only for the selected clones. In all cases, LacZ activity was measured in a quantitative luminometric assay (Tropix).

Identification of interacting fragments

The prey fragments of the positive clones were amplified by PCR, analysed on agarose gel, partially sequenced at their 5' junction on a PE 3700 Sequencer and mapped on the genomic sequence. Many clones were sequenced at their 3' junction to map precisely the SID. All the steps after picking of positive clones were performed in bar-coded 96-well plates and automated with Beckman Biomek 2000 and Multimek automats. At the end of each screening experiment, the identity of the bait plasmid was controlled on a few positive clones.

PIM biological score

The PIM biological score (PBS) computation relies on two different levels of analysis: first, a local (that is, taking into account only the results of one screen) score is computed for each screen; and second, the global score is computed from the local scores by integrating results from all screens performed within the same genomic library. Local scores are thus computed only once, while global scores are recomputed each time new screens are performed. For each screen, fragments are clustered by overlap to delimit SIDs. Fragments that have no or very improbable coding capability (antisense, intergenic region, and out-of-frame fusion fragments selected in a single frame) are then eliminated from the set of prey fragments identified from positive clones. Assuming that prey fragments compete for the bait with 'equal chances', the probability *p* for a given fragment to be selected in an experiment is proportional to its expected number of occurrences within the library. *p* is

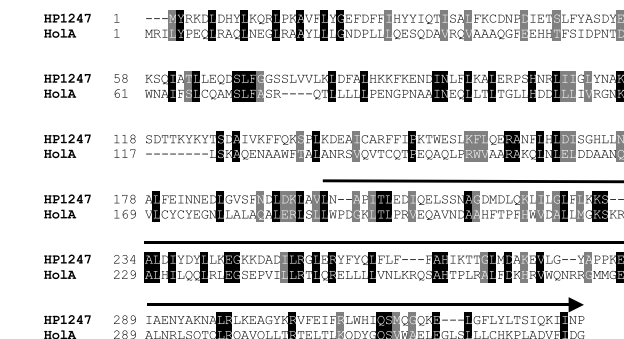


Figure 4 Alignment between *H. pylori* HP1247 protein and *E. coli* HoiA. The two sequences were aligned using the FastA algorithm. Identical (black) and similar (grey) residues are outlined. The position of HP1247 C-terminal domain interacting with HoiB is indicated by a line.

Table 2 Interacting proteins encoded by physically related genes

Protein 1		Protein 2	
ORF	Comments	ORF	Comments
HP0047†	HypE*, hydrogenase functionally related protein	HP0048	HypF*, hydrogenase putative transcriptional regulator
HP0061	–	HP0066	ATP-binding motif
HP0064	–	HP0063	–
HP0311	–	HP0312	ATP-binding motif
HP0338	–	HP0337	–
HP0691	ScoA, 3-oxoadipate coA-transferase subunit A	HP0692	ScoB, 3-oxoadipate coA-transferase subunit B
HP0697	–	HP0696	Putative hydantoin utilization*
HP0800	MoaE*, molybdopterin converting factor, subunit 2	HP0801	MoaD*, molybdopterin converting factor, subunit 1
HP0868	–	HP0869	HypA*, hydrogenase functionally related protein
HP0874	–	HP0875	KatA, catalase
HP0978	FtsA*, cell-division protein	HP0979	FtsZ*, cell-division protein
HP1244	RpsR, ribosomal protein	HP1246	RpsF, ribosomal protein

* Function assigned by homology.

† The nomenclature used is that of PyloriGene database.

computed as a function of the fragment length and position, and of the length and position distributions of fragments in the prey library (these distributions are calibrated using data from random sequencing).

The local score is the probability for a given SID to be obtained under the equal chance hypothesis, that is, as a result of random noise. It is deduced by combining probabilities p (using a binomial law) from each of the independent fragment defining it. A (global) PBS is computed for each protein interaction after pooling results from all screens. On the basis of an independence hypothesis, scores from different screens are combined together when the same protein domain pair is involved. The resulting PBS thus represents the probability that the protein–protein interaction is due to noise. Scores are real numbers ranging from 0 to 1, but are grouped in four categories (A, B, C and D) for practical purposes. Finally, the global connectivity of the interaction map is analysed to tag separately (category E) SIDs found as prey with frequency greater than a fixed threshold: the PBS of each protein–protein interaction involving highly connected SIDs is set to 1. Both the intercategory thresholds and the high-connectivity threshold were defined manually, taking into account the nature of the studied organism, the relevant library and the current coverage of the proteome ($A < 1^{e-10} < B < 1^{e-5} < C < 1^{e-2.5} < D$; the E category corresponds to prey SIDs selected with more than 4 baits and was arbitrarily attributed a PBS value of 1).

Bioinformatics

Several algorithms and software were implemented in the production database to facilitate experimental steps, such as a 'bait program' that designed automatically oligonucleotides for PCR amplification and sequencing of bait constructs, a 'prey program' that determined the position of each fragment in the genome and its coding capacity (such as intergene, antisense, nucleotide position in an ORF, coding frame). The interactions were then analysed through a web-based software platform, the PIMRider developed at Hybrigenics and accessible through the web interface (<http://pim.hybrigenics.com>). Academic users will be granted a free licence. Other users will have to purchase a commercial licence. The *H. pylori* PIMRider platform is linked to the PyloriGene database.

Received 8 August; accepted 10 October 2000.

- Fields, S. The future is function. *Nature Genet.* **15**, 325–327 (1997).
- Fromont-Racine, M., Rain, J. C. & Legrain, P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genet.* **16**, 277–282 (1997).
- Bartel, P. L., Roecklein, J. A., SenGupta, D. & Fields, S. A protein linkage map of *Escherichia coli* bacteriophage T7. *Nature Genet.* **12**, 72–77 (1996).
- Flajolet, M. *et al.* A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* **242**, 369–379 (2000).
- McCraith, S., Holtzman, T., Moss, B. & Fields, S. Genome-wide analysis of vaccinia virus protein–protein interactions. *Proc. Natl Acad. Sci. USA* **97**, 4879–4884 (2000).
- Ito, T. *et al.* Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA* **97**, 1143–1147 (2000).
- Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
- Walhout, A. J. M. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
- Tomb, J. F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547 (1997).
- Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
- Alm, R. A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180 (1999).
- Moszer, I. The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis. *FEBS Lett.* **430**, 28–36 (1998).
- Welch, M., Chinardet, N., Mourey, L., Birck, C. & Samama, J. P. Structure of the CheY-binding domain of histidine kinase CheA in complex with CheY. *Nature Struct. Biol.* **5**, 25–29 (1998).
- Cussac, V., Ferrero, R. L. & Labigne, A. Expression of *Helicobacter pylori* urease genes in *Escherichia coli* grown under nitrogen-limiting conditions. *J. Bacteriol.* **174**, 2466–2473 (1992).
- Mobley, H. L., Island, M. D. & Hausinger, R. P. Molecular biology of microbial ureases. *Microbiol. Rev.* **59**, 451–480 (1995).
- Skouloubris, S., Thiberge, J. M., Labigne, A. & De Reuse, H. The *Helicobacter pylori* UreI protein is not involved in urease activity but is essential for bacterial survival *in vivo*. *Infect. Immun.* **66**, 4517–4521 (1998).
- Weeks, D. L., Eskandari, S., Scott, D. R. & Sachs, G. A H⁺-gated urea channel: the link between *Helicobacter pylori* urease and gastric colonization. *Science* **287**, 482–485 (2000).
- Dong, Z., Onrust, R., Skangalis, M. & O'Donnell, M. DNA polymerase III accessory proteins. I. *hda* and *holB* encoding delta and delta'. *J. Biol. Chem.* **268**, 11758–11765 (1993).
- Liu, X. & Matsumura, P. An alternative sigma factor controls transcription of flagellar class-III operons in *Escherichia coli*: gene sequence, overproduction, purification and characterization. *Gene* **164**, 81–84 (1995).
- Zhang, G. *et al.* Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell* **98**, 811–824 (1999).
- Mooney, R. A. & Landick, R. RNA polymerase unveiled. *Cell* **98**, 687–690 (1999).
- Vidal, M. & Legrain, P. Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res.* **27**, 919–929 (1999).
- Ferrero, R. L., Cussac, V., Courcoux, P. & Labigne, A. Construction of isogenic urease-negative mutants of *Helicobacter pylori* by allelic exchange. *J. Bacteriol.* **174**, 4212–4217 (1992).

Acknowledgements

We thank M. Fromont-Racine, P. Glaser, A. Jacquier, A. Brunet and L. Decourty for their

help at the launch of this project; M. Fejes, G. Conan and P. Desmoucelle for technical assistance; G. Boissy and J.-L. Divol for their help in software development; F. Colland for his contribution to the mapping of FlhA interacting domain on the 3D structure of the core RNA polymerase; and S. Whiteside for a thorough and critical reading of the manuscript. We are very grateful to R. Benarous, J. Camonis, L. Daviet, M. Rosbash, A.D. Strosberg and S. Whiteside for many stimulating discussions. This work was supported by an interest-free loan from the ANVAR. P.L. is on leave from the CNRS.

Correspondence and requests for materials should be addressed to P.L. (e-mail: plegrain@hybrigenics.fr).

Crystal structures of SarA, a pleiotropic regulator of virulence genes in *S. aureus*

Maria A. Schumacher^{*†}, Barry K. Hurlburt[‡] & Richard G. Brennan^{*}

^{*} Department of Biochemistry and Molecular Biology and [†] Vollum Institute, Oregon Health Sciences University, Portland, Oregon 97201-3098, USA

[‡] Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, Arkansas 72205, USA

Staphylococcus aureus is a major human pathogen, the potency of which can be attributed to the regulated expression of an impressive array of virulence determinants. A key pleiotropic transcriptional regulator of these virulence factors is SarA, which is encoded by the *sar* (staphylococcal accessory regulator) locus^{1–3}. SarA was characterized initially as an activator of a second virulence regulatory locus, *agr*, through its interaction with a series of heptad repeats (AGTTAAG) within the *agr* promoter⁴. Subsequent DNA-binding studies have revealed that SarA binds readily to multiple AT-rich sequences of variable lengths^{4–11}. Here we describe the crystal structure of SarA and a SarA–DNA complex at resolutions of 2.50 Å and 2.95 Å, respectively. SarA has a fold consisting of a four-helix core region and 'inducible

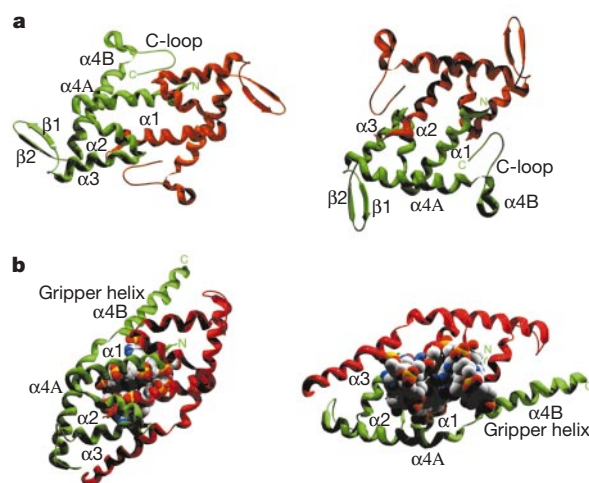


Figure 1 The SarA dimer and SarA–DNA complex. **a**, Views of SarA looking directly into the DNA-binding pocket or from the 'back' side. The secondary structures, β -hairpin and C-terminal loop of one monomer are labelled. Each monomer is red or green. **b**, The SarA–DNA complex in the identical orientation of the corresponding apo SarA directly above. The DNA duplex is shown as CPK atoms, with carbon, nitrogen, oxygen and phosphates coloured white, blue, red and yellow, respectively. The $\alpha 4B$ gripper helix of one monomer is also labelled. The narrow and deep minor groove and major groove are seen on the left and right respectively.

regions' comprising a β -hairpin and a carboxy-terminal loop. On binding DNA, the inducible regions undergo marked conformational changes, becoming part of extended and distorted α -helices, which encase the DNA. SarA recognizes an AT-rich site in which the DNA is highly overwound and adopts a D-DNA-like conformation by indirect readout. These structures thus provide insight into SarA-mediated transcription regulation.

We determined the crystal structure of full-length SarA (125 residues) from *S. aureus* strain DB by multiple-wavelength anomalous dispersion (MAD). Notably, on cryo-protection and freezing, crystals of SarA changed space group from $P2_12_12_1$ to $P2_12_12$ (Table 1), and the diffraction limit increased from about 3.0 Å to beyond 2.5 Å. The structure of SarA (Fig. 1a) has a new protein fold as ascertained by searches with the DALI server¹². The SarA monomer contains five α -helices, a short β -hairpin and a long C-terminal loop (C-loop). The topology is $\alpha 1$, (residues 10–30); $\alpha 2$, (residues 32–40); $\alpha 3$, (residues 44–56); $\beta 1$, (residues 62–64); $\beta 2$, (residues 68–70); $\alpha 4A$, (residues 74–94); $\alpha 4B$, (residues 97–103); and the C-loop, (residues 104–121). Two notable features are the protrusion of the β -hairpin and the C-loop from opposite ends of the monomer and the irregular conformation of helix $\alpha 4B$, which immediately precedes the C-loop (Fig. 1a). We only observed weak density for residues 116–121, which are located at the very end of the C-loop. These residues are not present in functional SarA protein

from some *S. aureus* strains, suggesting that they are dispensable for function¹³.

The physiologically relevant SarA dimer¹¹ is readily identified in the structure (Fig. 1a). The formation of this dimer leads to the creation of a hydrophobic core forged mainly from extensive contacts between $\alpha 1$ and $\alpha 1'$ (where the prime indicates the other subunit), which pack in an antiparallel fashion. This extensive dimerization interface buries 720 Å² of accessible surface area per monomer and remains essentially unchanged in the protein–DNA complex. The contribution of Phe80 to the nonpolar core of the dimer is notable, as it is the sole hydrophobic residue found on $\alpha 4A$, which provides a critical hydrophobic attachment point for anchoring this helix to the rest of the protein. In addition to this contact, a salt bridge between residues Glu 29 from $\alpha 1$ and Lys 83 from $\alpha 4A$ provides a second contact point between $\alpha 4A$ and the body of the protein.

The β -hairpin and C-loop, which project out from opposite ends of the protein, are tethered to the core by few contacts (Fig. 1a). The C-loop is anchored by residues 111–116, which loop back and pack against the amino terminus and residues from $\alpha 1$ of the same subunit. The β -hairpin is fastened to the body of the protein by stacking interactions between β -hairpin residue Tyr 62 and core residues, Phe 30 and Phe 80, and a hydrogen bond between β -hairpin residue Gln 64 and Tyr 92 from $\alpha 4A$. The main stabilization of these extended structures, however, is from crystal packing contacts in which the β -hairpin of one subunit packs against the C-loop of another subunit in the crystal. Such ordering is consistent with the marked increase in diffraction observed when data are collected under cryo-conditions, which may serve to 'freeze out' a specific conformation.

Examination of the SarA structure reveals no known DNA-binding motif or obvious DNA-binding pocket, with the possible exception of an unusual cleft that is formed at the dimer interface. This shallow cavity is formed by the $\alpha 1$ and $\alpha 1'$ helices, which would form the base, and the $\alpha 4A$ and $\alpha 4A'$ helices, which would form the sides (Fig. 1a). However, attempts to model a fragment of duplex B-DNA into this cleft were unsuccessful as the cleft is too narrow. We therefore used co-crystallization studies to determine the mechanism of DNA binding by SarA. Because there is no known SarA DNA consensus site, we carried out co-crystallization trials of SarA with numerous deoxynucleotides that encompassed the AT-

Table 1 SarA and SarA-DNA crystallographic data

SarA				
	$\lambda 1$	$\lambda 2$	$\lambda 3$	$\lambda 4$
λ	0.9798	0.9795	0.9250	1.0630
No. of total reflections	22,932	22,878	23,121	21,507
No. of unique reflections	5,757	5,763	5,740	5,660
Per cent complete	90.4	90.5	90.0	89.1
R_{sym} (%) [*]	8.5	8.7	8.7	8.2
$I/\sigma(I)$	4.7	5.0	4.9	4.9
Phasing power [†]	1.84/1.14	1.82/2.34	0.43/1.69	
R_{cullis} [‡]	0.605	0.625	0.798	
Figure of merit [§]	0.688			
Refinement				
Resolution limit (Å)	2.5			
R/R_{free} (%)	19.9/26.7			
r.m.s. deviation				
Bond angles (°)	2.10			
Bond lengths (Å)	0.013			
B values (Å ²)	4.09			

SarA-DNA				
	Native	Thimerosal		
Resolution	2.95	3.5		
No. of total reflections	14,850	10,297		
No. of unique reflections	10,951	8,623		
Per cent complete	89	80		
R_{sym} (%) [*]	7.8	10.8		
$I/\sigma(I)$	6.5	4.5		
Phasing power [†]		1.99		
R_{cullis} [‡]		0.505		
§ Figure of merit		0.408		
Refinement				
Resolution range	10.0-2.95			
R/R_{free} (%)	22.0/30.8			
r.m.s. deviation				
Bond angles (°)	1.98			
Bond lengths (Å)	0.011			
B values (Å ²)	4.12			

^{*} $R_{\text{sym}} = \sum |I_o - \langle I \rangle| / \sum I_o$, where I_o is the observed intensity and $\langle I \rangle$ is the average intensity obtained from multiple observations of symmetry related reflections.

[†] Phasing power = $r.m.s.(|F_o|/E)$, where $|F_o|$ is the heavy atom structure factor amplitude and E is residual lack of closure error.

[‡] $R_{\text{cullis}} = \sum ||F_o| - |F_{\text{calc}}|| / \sum |F_o|$ for centric reflections where $|F_o|$ is the observed heavy atom structure factor amplitudes and $|F_{\text{calc}}|$ is the calculated heavy atom structure factor amplitude.

[§] Figure of merit = $\langle |\Sigma P(\alpha)e^{i\alpha} / \Sigma P(\alpha)| \rangle$, where α is the phase and $P(\alpha)$ is the phase probability distribution.

^{||} $R = \sum ||F_o| - |F_{\text{calc}}|| / \sum |F_o|$. The lower resolution limit of refinement for all structures was 10.0 Å and $R_{\text{free}} = \sum ||F_o| - |F_{\text{calc}}|| / \sum |F_o|$, where all reflections belong to a test set of 10% randomly selected data.

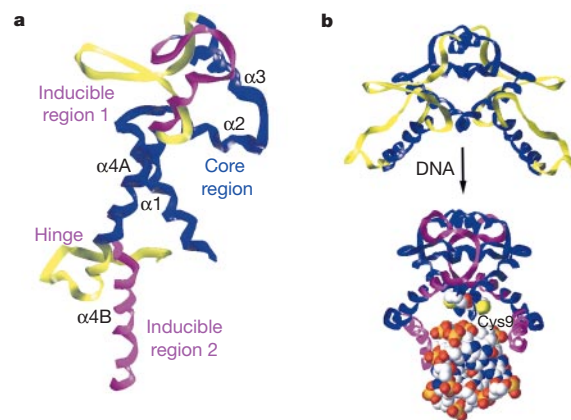


Figure 2 SarA inducible regions. **a**, Superimposed SarA monomers of the DNA-bound and apo protein. The core region is blue and the inducible regions are yellow and magenta for the apo and DNA bound proteins, respectively. The hinge is labelled. **b**, Conformational changes induced by DNA binding. The apo form is shown on top (inducible regions in yellow) and the DNA-bound form below (inducible regions in magenta). The DNA duplex and Cys 9 side chains, which line the DNA binding channel, are shown as CPK atoms, with carbon, nitrogen, oxygen and phosphate/sulphur coloured white, blue, red and yellow, respectively.

rich heptad repeat (AGTTAAG) located within the *agr* promoter⁴. Synthetic DNA containing multiple heptad repeats was shown to bind SarA and could be used to purify the protein from *S. aureus* extracts⁹. Because the minimal SarA-binding site is unclear, we used deoxyoligonucleotides duplexes ranging from 7 to 42 bp (base pairs). Data quality crystals were obtained only with an 8-bp duplex (8-mer) with 5'A and T overhangs, (5'AGTTAAGA3')-(3'CAATTCTT5').

We determined the structure of the full-length SarA–8-mer complex by single isomorphous replacement, which uses phases from a thimerosal derivatization of the highly reactive cysteine, Cys 9 (Table 1). Individual substitution of each thymine with 5-iodouracil failed to reveal any iodine sites, indicating that the DNA was disordered. Despite this disorder, clear electron density was observed for the phosphodiester backbone of 6 bp and the 3' nucleotide at the end of each strand, as well as the SarA dimer (Fig. 1b).

Comparison of the SarA structure of the DNA bound form with that of the apo form shows that the four-helix core region, $\alpha 1$, $\alpha 2$, $\alpha 3$ and $\alpha 4A$, are essentially identical with root mean square (r.m.s.) deviations of 0.6 Å for the corresponding C α atoms of residues 10–45 and 76–90 (Fig. 2a). However, the extended regions comprised of the β -hairpin and C-loop regions, which lie on either side of $\alpha 4A$, have undergone marked conformational changes. Because of their obvious structural flexibility, we called these the 'inducible regions' (Fig. 2a). Structural changes in inducible region 1 include the conversions of $\beta 2$ of the β -hairpin and five additional residues (71–75) to a helical conformation, which extend $\alpha 4A$ by three turns at its N terminus (Fig. 2a, b). Structural changes in inducible region 2 involve the incorporation of residues 103–114 of the C-loop and residues 94–95 into helix $\alpha 4B$, which increases to six full turns. The net result of these large conformational changes is the formation of an extended and distorted α -helix 4 (residues 67–114), which thoroughly encases the DNA. The main distortion in $\alpha 4$ occurs at residues 91–93, which seem to act as a flexible hinge that allows the $\alpha 4B$ helices to re-orient and encase the DNA, a process impossible with a single, long helix. Such flexibility is consistent with the different orientations of these 'gripper helices' in each monomer, whereby one is more proximal and the other more distal to the DNA (Fig. 1b). The inducible regions seem to undergo their structural transitions to avoid C-loop–DNA steric clash and to maximize electrostatic and hydrophobic interactions with the DNA.

The DNA is bound in the long channel between $\alpha 4A$ and $\alpha 4A'$ and the helices of the inducible regions of the SarA dimer (Figs 1b and 2b). As noted, this pocket is too narrow to accommodate a duplex B-DNA structure. Instead the DNA adopts an unusual conformation that is highly overwound and marked by a very

narrow minor groove. In contrast to canonical B-DNA, which has an average helical twist of 34.3°, a minor-groove width of 5.7 Å and 10.5 bp per turn, the SarA-bound DNA displays an average helical twist of 41.8°, a minor-groove width of 2.8 Å and 8.6 bp per turn (Fig. 3)¹⁴. Notably, there is a similar DNA conformation in fibre diffraction studies on DNA with AT and AATT repeats^{15,16}. This DNA, called D-DNA, is also characterized by an extremely narrow minor groove (1.7 Å), a large helical twist of 44.8° and an overwound helix with 8 bp per turn (Fig. 3). Fibre diffraction studies on such AT-rich repeats show that these polymers do not adopt or transit through the A-DNA conformation, and polymers of (dA-dA-dT)•(dA-dT-dT) adopt only a B-DNA or D-DNA conformation depending on the environment. Specifically, divalent cations and low humidity favour D-DNA¹⁵.

The 'relaxed sequence' binding of SarA is consistent with the lack of base-specific contacts observed in the SarA–DNA complex, and suggests that indirect readout is the primary mode of DNA recognition. SarA is anchored to the DNA by phosphate contacts of residues from helices $\alpha 1/\alpha 1'$ and $\alpha 4A/\alpha 4A'$. All contacts are made to the minor groove of the DNA, which faces the protein. A key set of SarA–phosphate contacts is provided by an extended network of hydrogen bonds and ionic interactions at either end of the dimer dyad consisting of P₅–Tyr 18/Arg 84–Glu 11'–Ca²⁺–P_{7'} (Fig. 4). This network both requires and buttresses the narrow minor groove, which is an intrinsic property of the D-DNA conformation (Fig. 4)¹⁵.

In addition to these phosphate backbone contacts, SarA makes a number of hydrophobic contacts to the DNA minor groove. These interactions are made possible by the structural changes that allow SarA to encase the DNA and which result in the creation of an environment favouring D-DNA formation¹⁵. Specifically, there is a

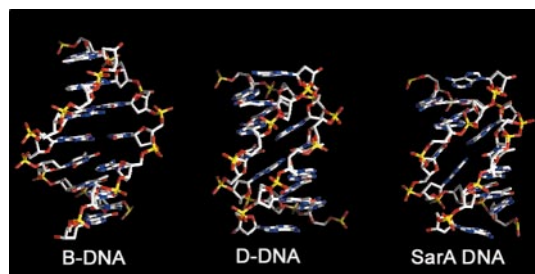


Figure 3 Conformation of SarA-bound DNA. B-DNA, D-DNA and DNA from the SarA–DNA complex are shown. Each 6-bp fragment is shown in the same orientation. We note the extremely narrow minor groove and overwound helices of the D-DNA and SarA-bound DNA structures, which are in sharp contrast to those B-DNA conformational features. Carbon, nitrogen, oxygen and phosphorous atoms are coloured white, blue, red and yellow, respectively.

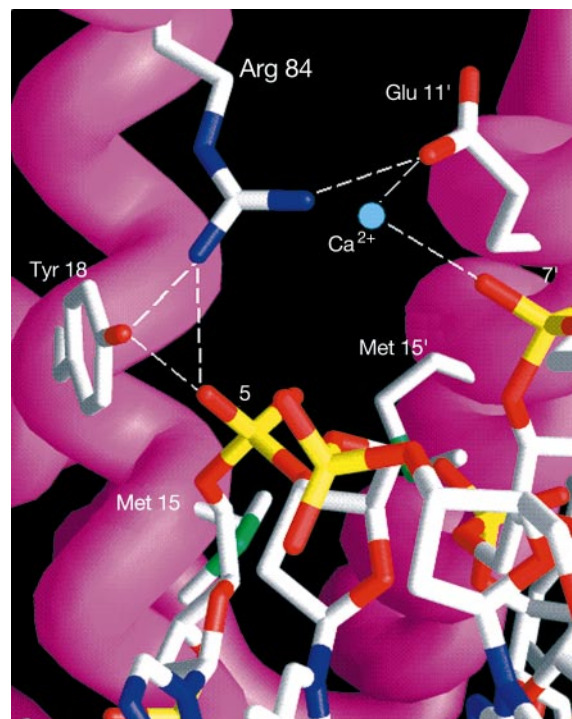


Figure 4 SarA–DNA contacts. One end of the network of protein–DNA interactions that stabilize the narrow minor groove. The side chains of Tyr 18, Met 15, Arg 84, Glu 11' and Met 15' and the DNA are shown as sticks; carbon (white), oxygen (red), nitrogen (blue), sulphur (green) and phosphorus (yellow) atoms are also shown. The calcium ion is depicted as a light blue sphere and interactions are shown as dashed lines. The six base pairs and 3' nucleotide overhangs of the DNA duplex are labelled as in Methods. Figures 3 and 4 were made with GRASP²⁹.

large nonpolar face at the centre of the DNA-binding site created by residues Met 15 and Met 15'. These residues abut directly and interact with the deoxyribose phosphate backbone of base pairs 5 and 6 and 5' and 6', respectively (Fig. 4). Furthermore, the $\alpha 4A/\alpha 4B$ hinge and the N terminus of $\alpha 4B$, including residues Val 92, Leu 93, Ile 94, Leu 95 and Val 96, contribute to the hydrophobic environment of the binding pocket and may be potentially involved in van der Waals contacts to the bases. Indeed, the side chain of Leu 95 from the DNA proximal gripper helix is close to Ade 5. More extensive contacts with these aliphatic residues would require a longer DNA site that is bent towards the protein.

Other residues of the $\alpha 4B$ gripper helices are also likely to be involved in binding longer DNA sites, which, given their ability to adopt several orientations, could wrap around these sites. These helices contain a highly positively charged face, comprising Lys 102, Lys 103, Arg 110, Lys 113 and Arg 114, and their two-fold equivalents, which line the far ends of the DNA-binding cleft and make long-range electrostatic contacts to the DNA. The $\alpha 4B$ helix is highly analogous in structure, dynamics and function to the DNA-binding regions of the basic leucine zipper and basic helix-loop-helix proteins, as all three undergo a coil to helix transition on DNA binding, and contain numerous basic residues involved, or probably involved, in DNA binding¹⁷. The structures of SarA and the SarA-DNA complex seem to have captured for the first time the

conformational end states of a DNA-induced coil-to-helix folding transition.

SarA both represses and activates transcription from different promoters. Moreover, activation of the *agr* locus is only partially responsible for the global effect of SarA on virulence gene expression, as both *sar*-dependent repression of genes encoding surface proteins and *sar*-dependent activation of genes encoding exoproteins occur⁵⁻⁸. What could explain the basis for such an apparent dichotomy in function? The SarA-DNA structure shows the marked overwinding of the DNA binding site that results in a significant decrease in the number of base pairs per turn of helix. This altered conformation could directly explain the observation that SarA activates the *agr* P3 promoter, in which there are about 20 bp between the -35 and -10 boxes⁴, and represses transcription of the collagen-binding protein gene (*cna*), where the 17-bp spacer between -35 and -10 is optimal⁷. In the former site, the overwinding would reduce the -35/-10 spacer to 17 bp, and in the latter the 'optimal' spacer would shrink to 14 bp. However, the overall regulatory effects of SarA are likely to be more complex, as indicated by the ability of this protein to bind numerous DNA sites in the *S. aureus* genome⁴⁻¹¹. SarA, like the *Escherichia coli* proteins Fis, IHF and HU, might therefore function as a global, architectural DNA-binding protein that influences DNA superhelicity and, in turn, transcription in a manner dependent on binding-site position¹⁸⁻²⁰.

SarA regulates virulence genes in a growth-phase-dependent manner, a feature critical to the progression of infection in *S. aureus*. Therefore, the temporal regulation of SarA binding becomes critical. In this regard, we note that the *sar* locus encodes three overlapping transcripts all of which encode the SarA protein. Although these transcripts are initiated by different σ -factors at different times in the *S. aureus* life cycle, the total amount of SarA protein seems to remain constant⁷. This suggests that the activity of SarA is regulated by either its binding to other proteins or its post-translational modification. In support of the latter hypothesis, SarA activity is significantly reduced under oxygen-replete conditions²¹. This is particularly intriguing because Cys 9, the single cysteine in SarA, is highly reactive and is located in the DNA binding channel (Fig. 2b). Oxidation of the thiol side chain to a sulphone or cysteic acid would interfere directly with DNA binding.

Given the structural uniqueness and critical function of SarA in *S. aureus* virulence, the structures described here provide scaffolds from which to initiate the design of new and highly specific anti-staphylococcal chemotherapeutics. □

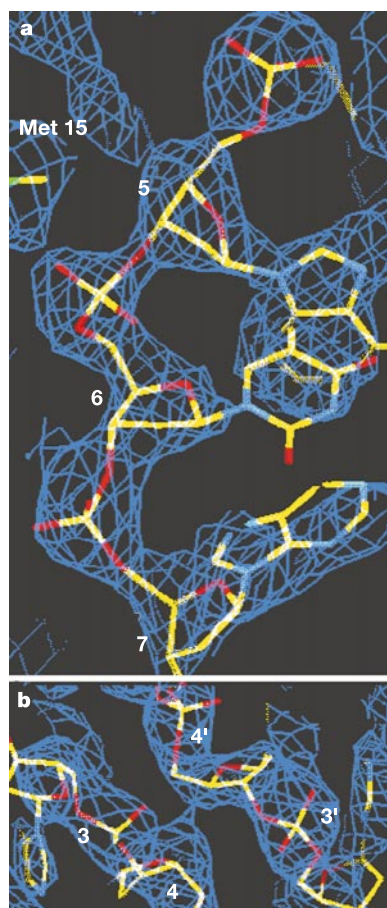


Figure 5 $2F_{\text{obs}} - F_{\text{calc}}$ omit electron density maps in which the DNA has been omitted. **a**, View of the omitted electron density map showing the region about the bases and phosphodiester backbone of nucleotides 5-7. **b**, View of the omitted electron density map highlighting the very close approach of the phosphodiester backbones across the minor groove. Both **a** and **b** are contoured at 1.2σ . The DNA is depicted as sticks, in which carbon, nitrogen, oxygen and phosphorous atoms are coloured yellow, blue, red and orange, respectively.

Methods

SarA purification and crystallization

We purified full-length recombinant SarA from strain DB as described¹¹. We grew 5231 of bacteria to obtain enough protein to determine the structures of SarA and the SarA-DNA complex. We grew crystals of SarA by hanging drop/vapour diffusion at 4 °C using the protein at 10 mg ml⁻¹ and a precipitant of 6% isopropanol, in 50 mM Tris, pH 7.6 and 10 mM MgCl₂. We characterized the space group at room temperature as orthorhombic, $P2_12_12_1$ with $a = 44.8 \text{ \AA}$, $b = 84.4 \text{ \AA}$, $c = 141 \text{ \AA}$; however, at room temperature the crystals decay rapidly. Only glycerol worked as a cryoprotectant, yet typically produced a large increase in mosaic spread (0.3° to $\sim 2.2^\circ$). Cryoprotection extends the diffraction limit from 3.0 Å to beyond 2.5 Å resolution and, notably, changes the space group to $P2_12_12$ with $a = 44.8 \text{ \AA}$, $b = 84.4 \text{ \AA}$, $c = 27.6 \text{ \AA}$. The cryoprotected crystals contain a monomer per asymmetric unit and a solvent content of 25%.

Structure determination and refinement

We determined the structure by MAD²², as individual frozen crystals display severe nonisomorphism. Because the wild-type protein contains only one methionine, a mutant, L95M, was constructed and selenomethionine L95M SarA was purified and isomorphously crystallized. We collected MAD data at the Stanford synchrotron radiation laboratory at line 1-5 using an ADSC 180 mm charged coupled device (CCD) detector. We processed the data with MOSFLM²³. Only one site, corresponding to Met 95, was located in Patterson maps; the second site was found by difference Fourier techniques.

We refined the selenium atomic positions by treating the data as a special case of MIR²⁴. The resulting map revealed clear density for helices and the β -hairpin. Model building²⁵ was aided by phase combination. We refined the structure with TNT²⁶. Although the data extend beyond 2.5 Å resolution, the refinement was limited to 2.5 Å owing to problems

with the crystal mosaicity and anisotropy of the data. The current structure consists of residues 2–121 and 82 waters with 92% of the residues in favoured regions of the Ramachandran plot, 6% in generously allowed regions and 2% in unfavoured regions²⁷. Residues Ile 104 and Ala 118 are the two outliers; Ile 104 lies in the tight turn between $\alpha 4B$ and the C-loop; and Ala 118 has weak density along with residues 116–121

Crystallization of the SarA–DNA complex

We used a number of deoxyoligonucleotides in crystallization trials with SarA, and only those with an 8-mer duplex yielded data-quality crystals. We grew these crystals using hanging drop/vapour diffusion with a crystallization solution of 25% PEG 8000, 200 mM CaCl₂, 100 mM cacodylate, pH 6.5. They are monoclinic, space group P2₁, with $a = 54.5$ Å, $b = 65.2$ Å, $c = 57.8$ Å and $\beta = 118.0^\circ$ and contain one SarA dimer and 8-mer duplex per asymmetric unit. All data were collected at room temperature with an ADSC multiwire area detector²⁸ and a RIGAKU RU200-H rotating anode X-ray generator with graphite monochromator operating at 40 kV and 150 mA.

Structure determination and refinement

Because of DNA disorder, iodinated deoxyoligonucleotides failed as derivatives, therefore phases were derived from a thimerosal derivative in conjunction with twofold averaging and solvent flattening²⁴. Density was evident for the $\alpha 1$ and $\alpha 4A$ helices and the backbone of helices $\alpha 2$ and $\alpha 3$. However, in averaged maps, the density for $\alpha 4B$ was smeared and subsequently phase-combined; unaveraged maps revealed different orientations of these helices. The density for the phosphate backbone of 6 bp in the centre of the DNA-binding channel was also evident (Fig. 5). Good electron density was also observed for the bases but their identities were ambiguous and therefore, fitted with alternating A-T base pairs (5'-A₁T₂A₃T₄A₅T₆A₇-3'). We then subjected the structure to positional and tightly restrained B-factor refinement²⁶. The current model includes residues 2–113 of one subunit and 2–121 of the other, and a 6-bp duplex with 3'-adenylate nucleotide overhangs. Analysis of the Ramachandran plot shows that 94% of the residues are in the most favoured regions, 5.5% in generously allowed regions and 0.5% in unfavoured regions (Phe 30 from one subunit is the outlier)²⁷. Two large solvent peaks were located in the DNA minor-groove binding site. Their thermal parameters indicated that they are Ca²⁺ ions, which is consistent with the observation that either magnesium or calcium is required for crystallization of the SarA–DNA complex. Residual density in the DNA-binding channel indicates additional disordered DNA in the binding site. Although the gripper helices $\alpha 4B$ and $\alpha 4B'$ take different orientations, residues 5–95 of each subunit overlay with r.m.s. deviations of 0.55 Å.

Received 10 August; accepted 16 October 2000.

- Projan, S. L. & Novick, R. P. In *Staphylococci in Human Disease* (eds Cross, K. B. & Archer, G. L.) 55–81 (Churchill Livingstone, New York, 1997).
- Cheung, A. L., Koomey, J. M., Butler, C. A., Projan, S. J. & Fischetti, V. A. Regulation of exoprotein expression in *Staphylococcus aureus* by a locus (*sar*) distinct from *agr*. *Proc. Natl Acad. Sci. USA* **89**, 6462–6466 (1992).
- Cheung, A. L. & Projan, S. J. Cloning and sequencing of *sarA* of *Staphylococcus aureus*, a gene required for expression of *agr*. *J. Bacteriol.* **176**, 4168–4172 (1994).
- Morfeldt, E., Tegmark, K. & Arvidson, S. Transcriptional control of the *agr*-dependent virulence gene regulator, RNAIII, in *Staphylococcus aureus*. *Mol. Microbiol.* **21**, 1227–1237 (1996).
- Wolz, C. et al. *Agg*-independent regulation of fibronectin-binding protein(s) by the regulatory locus *sar* in *Staphylococcus aureus*. *Mol. Microbiol.* **36**, 230–243 (2000).
- Cheung, A. & Ying, P. Regulation of α - and β -hemolysins by the *sar* locus of *Staphylococcus aureus*. *J. Bacteriol.* **176**, 580–585 (1994).
- Blevins, J. S., Gillaspay, A. E., Reicht, T. M., Hurlbert, B. K. & Smeltzer, M. S. The staphylococcal accessory gene regulator (*sar*) represses transcription of the *Staphylococcus aureus* collagen adhesin gene (*cna*) in an *agr*-independent manner. *Mol. Microbiol.* **33**, 317–326 (1999).
- Gaskill, M. E. & Khan, S. A. Regulation of the enterotoxin B gene in *Staphylococcus aureus*. *J. Biol. Chem.* **263**, 6276–6280 (1988).
- Manna, A. C., Bayer, M. G. & Cheung, A. L. Transcriptional analysis of different promoters in the *sar* locus in *Staphylococcus aureus*. *J. Bacteriol.* **180**, 3828–3836 (1998).
- Chien, Y.-t., Manna, A., Projan, S. J. & Cheung, A. L. SarA, a global regulator of virulence determinants in *Staphylococcus aureus*, binds to a conserved motif essential for *sar*-dependent gene regulation. *J. Biol. Chem.* **274**, 37169–37176 (1999).
- Reicht, T. M. et al. Characterization of the SarA virulence gene regulator of *Staphylococcus aureus*. *Mol. Microbiol.* **33**, 307–316 (1999).
- Holm, L. & Sander, C. Protein structure comparisons by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
- Heinrichs, J. H., Bayer, M. G. & Cheung, A. L. Characterization of the *sar* locus and its interactions with *agr* in *Staphylococcus aureus*. *J. Bacteriol.* **178**, 418–423 (1996).
- Lavery, R. & Sklenar, H. J. The definition of generalized helical parameters and of axis curvature for irregular nucleic acids. *Biomol. Struct. Dynam.* **6**, 63–91 (1988).
- Saenger, W. *Principles of Nucleic Acid Structure* (Springer, New York, 1984).
- Mahendrasingam, A. et al. Time-resolved X-ray diffraction studies of the B $\leftrightarrow$ D structural transition in the DNA double helix. *Science* **233**, 195–197 (1986).
- Baxevanis, A. D. & Vinson, C. R. Interactions of coiled coils in transcription factors: where is the specificity? *Curr. Opin. Genet. Dev.* **3**, 278–285 (1993).
- Pruss, G. & Drlica, K. DNA supercoiling and prokaryotic transcription. *Cell* **56**, 521–523 (1989).
- Record, M. T. Jr, Reznikoff, W. S., Craig, M. L., McQuade, K. L. & Schlax, P. J. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* (eds Neidhardt, F. C. et al.) 792–820 (ASM, Washington DC, 1996).
- Oberto, J., Drlica, K. & Rouviere-Yaniv, J. Histones, HMG, HU, IHF: Meme combat. *Biochimie* **76**, 901–908 (1994).
- Chan, P. F. & Foster, S. J. Role of SarA in virulence determinant production and environmental signal transduction in *Staphylococcus aureus*. *J. Bacteriol.* **180**, 6232–6241 (1998).

- Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51–58 (1991).
- Collaborative computational project number 4. *J. Appl. Crystallogr.* **50**, 760 (1994).
- Furey, W. B. & Swaminathan, S. PHASE-95: a program package for the processing and analysis of diffraction data from macromolecules. In *Methods in Enzymol.* **Vol. 277** (eds Carter, C. W. & Sweet, R. M.) 590–620 (Academic, Orlando, 1997).
- Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta. Crystallogr. A* **47**, 110–119 (1991).
- Tronrud, D. E., TenEyck, L. F. & Matthews, B. W. An efficient general purpose least-squares refinement program for macromolecular structures. *Acta. Crystallogr. A* **43**, 489–501 (1985).
- Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291 (1993).
- Howard, A. J., Nielson, C. & Xuong, N. H. Software for a diffractometer with multiwire area detector. *Methods Enzymol.* **114**, 452–472 (1985).
- Nicholls, A., Sharp, K. & Honig, B. H. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).

Acknowledgements

We thank M. S. Smeltzer for his critical reading of this manuscript. Intensity data collected at the Stanford Synchrotron Radiation Laboratory (SSRL) was carried out under the SSRL biotechnology program, which is supported by the National Institutes of Health (NIH), National Center for Research Resources, Biomedical Technology Program, and by the Department of Energy, Office of Biological and Environmental Research. M.A.S. is a Burroughs Wellcome Career Awardee. This work was supported by the NIH Oregon Health Sciences Foundation (R.G.B.).

Correspondence and requests for materials should be addressed to R.G.B. (e-mail: brennanr@ohsu.edu). The refined coordinates for the SarA structure and SarA–DNA complex have been deposited with the Protein Data Bank under accession codes 1FZN and 1FZP, respectively.

Projection structure of a ClC-type chloride channel at 6.5 Å resolution

Joseph A. Mindell*, Merritt Maduke*, Christopher Miller* & Nikolaus Grigorieff†

* Department of Biochemistry, Howard Hughes Medical Institute and † Rosenstiel Basic Medical Science Research Center and W. M. Keck Institute for Cellular Visualization, Brandeis University, Waltham, Massachusetts 02454, USA

Virtually all cells in all eukaryotic organisms express ion channels of the ClC type, the only known molecular family of chloride-ion-selective channels. The diversity of ClC channels highlights the multitude and range of functions served by gated chloride-ion conduction in biological membranes, such as controlling electrical excitability in skeletal muscle, maintaining systemic blood pressure, acidifying endosomal compartments, and regulating electrical responses of GABA (γ -aminobutyric acid)-containing interneurons in the central nervous system¹. Previously, we expressed and purified a prokaryotic ClC channel homologue². Here we report the formation of two-dimensional crystals of this ClC channel protein reconstituted into phospholipid bilayer membranes. Cryo-electron microscopic analysis of these crystals yields a projection structure at 6.5 Å resolution, which shows off-axis water-filled pores within the dimeric channel complex.

Of the approximately ten ClC isoforms found in eukaryotes, the muscle subtypes ClC-0 and ClC-1 are by far the most thoroughly studied. These voltage-gated channels are homodimers of polypeptides with a relative molecular mass of 90,000 that contain 10–12 putative transmembrane segments³. Detailed single-channel analysis indicates that these channels are double-barrelled homodimers in which the functional dimer contains two identical pores devoid of axial symmetry, each with its own voltage-dependent gate^{4–8}. It is not known, however, whether this unusual molecular architecture applies to all ClC channels or only to the muscle subfamily. Moreover, this picture is not universally accepted; recently, a conventional

barrel-stave architecture for CIC-1 was proposed, with a single pore formed on the homodimer's two-fold axis⁹. Such fundamentally discordant views are best resolved with direct structural information; the recent overexpression, purification and functional reconstitution of EriC, a prokaryotic CIC channel², provides opportunities for structural approaches to the problem.

We produced two-dimensional crystals of EriC by reconstituting the purified channel protein into phospholipid membranes at high protein-to-lipid ratio. The crystals, which form spontaneously during removal of detergent by dialysis over several days, appear in large sheets up to 4 μm on a side (Fig. 1a) containing 10^4 – 10^5 EriC channels; most of the material in these samples is crystalline, as gauged by optical diffraction of negatively stained electron microscope images. Although double and multi-layer crystals are often observed, single-layered sheets are common, and only these were analysed here.

Before beginning structural analysis, we assessed whether the crystallization process irreversibly damages the protein. Crystalline EriC was solubilized in detergent and reconstituted back into lipid vesicles at much lower protein density, so that each vesicle carries approximately a single EriC channel. Efflux of $^{36}\text{Cl}^-$ from these vesicles was then assayed by a method designed to quantify the fraction of protein in the preparation that is functionally active². Figure 1b shows that EriC recovered from crystals is indistinguishable in activity from EriC that had never been crystallized. This

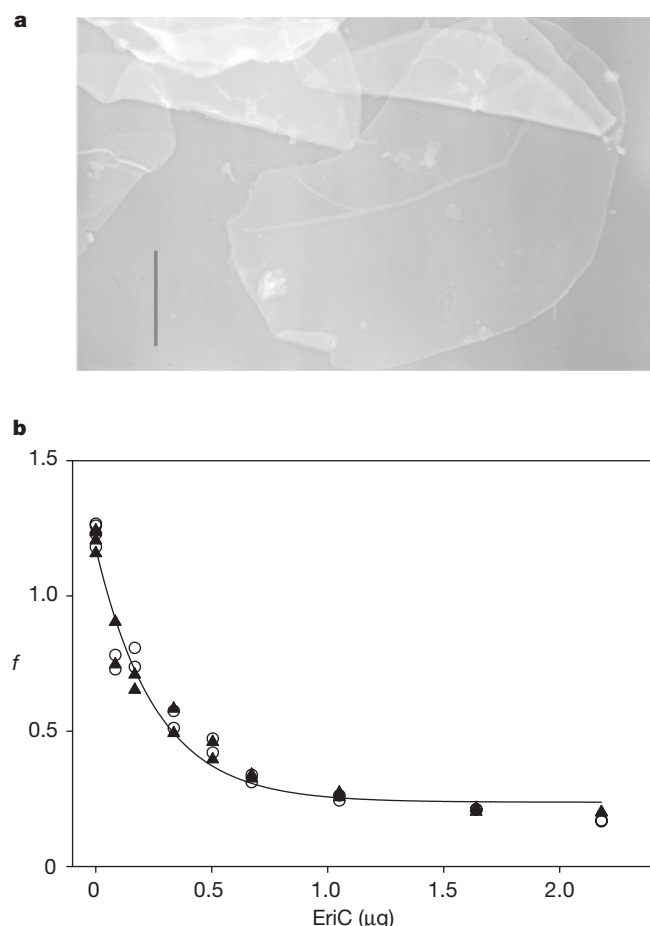


Figure 1 Active 2D crystals of EriC. **a**, Electron microscopic image of crystals embedded in 1% uranyl acetate stain. Scale bar, 1 μm . **b**, Functional activity of EriC after crystallization. The fractional trapped $^{36}\text{Cl}^-/{}^3\text{H}[\text{glutamate}]$ ratio, f , was determined for vesicles reconstituted with the indicated amount of EriC (closed triangles) or EriC extracted from the 2D crystals (open circles). The solid curve represents the variation expected from a Poisson distribution of protein in vesicles².

result gives us confidence as to the integrity of the protein in the crystals.

The computed diffraction pattern of unstained, glucose-embedded EriC crystals (Fig. 2a) reveals an orthorhombic unit cell with dimensions of $140 \times 80 \text{ \AA}$. Systematic absences of odd axial reflections suggest glide planes along both axes. These low-dose images show strong reflections to $\sim 8.5 \text{ \AA}$ by optical diffraction, and after correction for lattice distortions yield significant data to at least $5 \text{ \AA}$; in addition, electron diffraction produces reflections to $4 \text{ \AA}$ (data not shown). For all images, analysis of the phases of symmetry-related spots¹⁰ of both raw and processed transforms places the crystals in the two-sided plane group $P22_12_1$. Inspection of computationally filtered but unsymmetrized images (Fig. 2b) further confirms $P22_12_1$ symmetry, which arises from packing of near neighbours in opposite transmembrane orientations.

From about 80 micrographs, 7 were chosen for further processing. The crystalline areas of these images were corrected for lattice distortions, merged and averaged in $P22_12_1$. Crystallographic parameters are reported in Table 1, and Fig. 3 indicates the quality

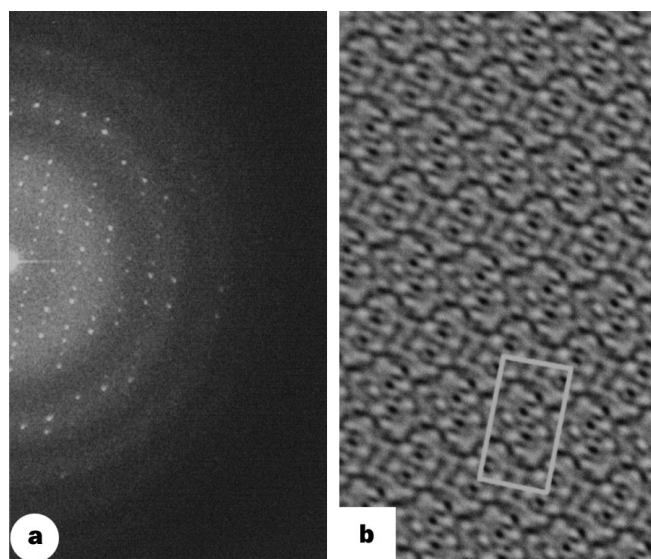


Figure 2 Analysis of glucose-embedded EriC crystals. **a**, Computed diffraction pattern from a single cryo-electron microscope image of an EriC crystal, showing strong spots out to $\sim 8.5 \text{ \AA}$ resolution. **b**, Digitally filtered image of EriC crystal. The diffraction pattern in **a** was filtered by removing all data not on the reciprocal lattice. This masked transform was then corrected for the CTF and reverse Fourier transformed to yield an image of the crystal. In this grey-scale image, higher scattering densities are brighter, and lower densities are darker. A single unit cell is boxed.

Table 1 Electron crystallographic image statistics

Two-sided plane group symmetry	$P22_12_1$	
Unit cell dimensions	$a = 140 \text{ \AA}$ $b = 80 \text{ \AA}$ $\gamma = 90^\circ$	
Number of images and lattices	7	
Range of defocus	4,280–8,390 $\text{\AA}$	
Total number of reflections	1,766	
Number of unique reflections	204	
Overall phase residual to $6.5 \text{ \AA}$ (random = 45°)	12.5°	
Resolution range ($\text{\AA}$)	Number of unique reflections	Phase residual (random = 45°)
100–10.01	84	4.2
10.0–8.01	51	9.9
8.0–7.01	42	22.1
7.0–6.51	27	28.4
6.5–6.01	36	36.1
6.0–5.0	94	40.8

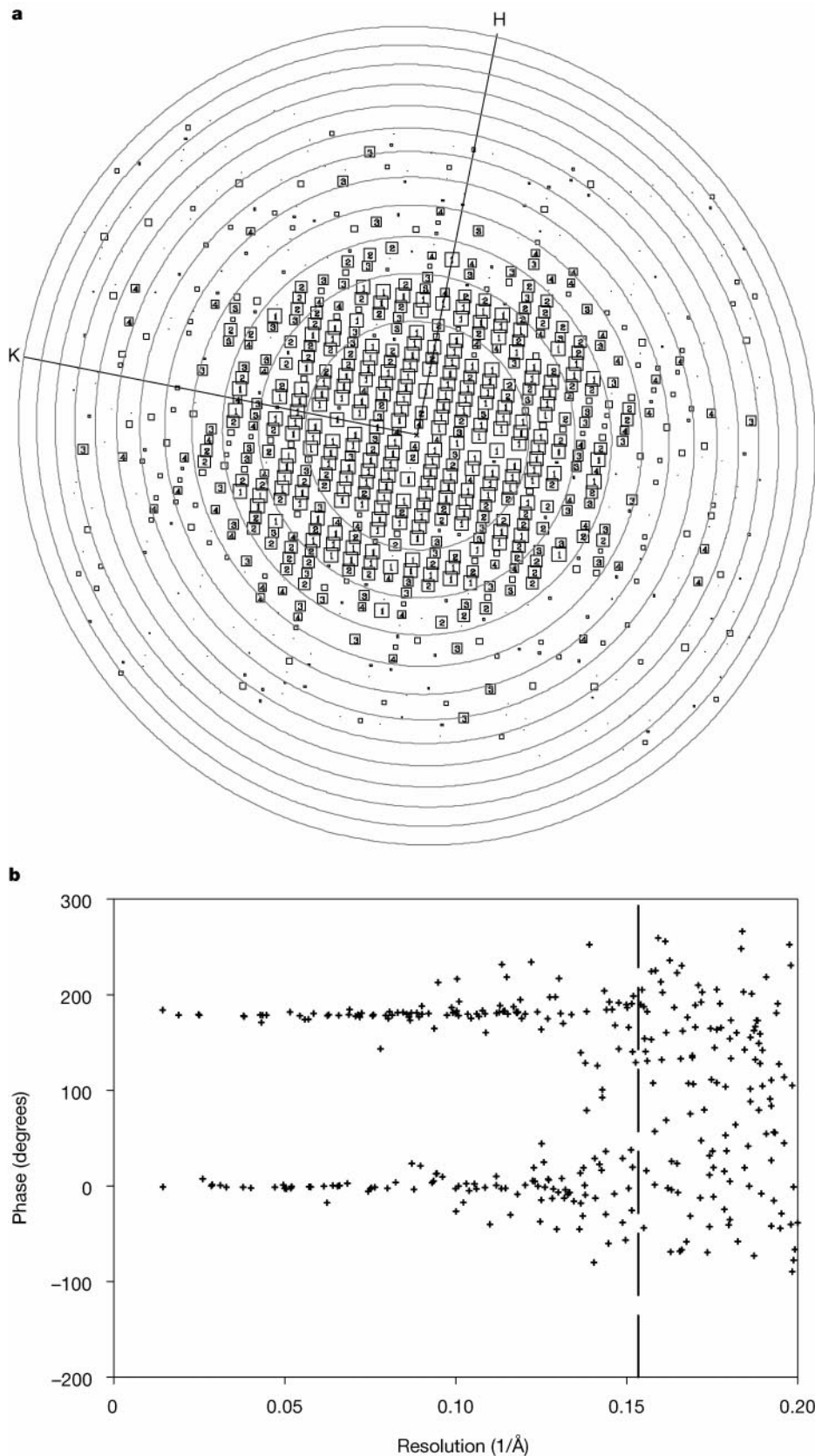


Figure 3 Quality of diffraction data. **a**, Plot of diffraction spots from the image described in Fig. 2, after correction for crystal distortion. Each spot in the transform is represented by a square and number (IQ value¹⁴) indicating the signal-to-background ratio (S/B), with the largest boxes and smallest numbers reflecting the best data (IQ = 1, S/B > 7.0; IQ = 2, S/B > 3.5; IQ = 3, S/B > 2.3; IQ = 4, S/B > 1.8; IQ = 5, S/B > 1.4; IQ = 6, S/B > 1.2;

IQ = 7, S/B > 1.0; IQ = 8, S/B > 0; IQ = 9, S/B < 0). Values from 1 to 4 are shown as numbers inside boxes, values from 5 to 8 are indicated by decreasing box size. The rings indicate zeros in the CTF. H and K indicate the reciprocal axes. **b**, Measured phases as a function of resolution. Each point represents the averaged phases for a given reflection from seven images. The cutoff for the data used here is indicated by a dashed line.

of reflections and measured phases as a function of resolution. Raw phases were rounded to 0° or 180° to impose strict two-fold symmetry on the projection, and phase residuals were calculated from the absolute values of the rounding errors. Although the averaged phase residuals are below random (45°) to 5.0° (Fig. 3), we used data only to $6.5\text{ \AA}$. In addition, only spots with a signal/background ratio ≥ 1.4 were used. In this merged dataset, the overall phase residual was 12.5° , with a 28.4° residual in the highest-resolution range used, $7\text{--}6.5\text{ \AA}$.

In projection normal to the membrane plane, EriC is a lens-shaped protein roughly $100\text{ \AA}$ in length and $50\text{ \AA}$ in width (Fig. 4a). A single unit cell contains two EriC channels; crystallographic two-fold axes lie within the protein and thereby confirm the expected homodimeric nature of the complex^{2,11}. The protein–lipid boundary is unmistakable, with the flat, featureless lipid areas distinct from the steeply contoured higher density of the protein. EriC shows two pairs of remarkably low-density troughs, $\sim 8\text{ \AA}$ and $\sim 5\text{ \AA}$ in width, each pair separated by a density peak $\sim 7\text{ \AA}$ across. These trough-densities are substantially lower than the lipid density

observed between the channel complexes (Fig. 4b).

Of the substances present on the sample grid—protein, lipid, water and glucose—the weakest scatterer is water¹². This fact alone identifies the low-density wells as aqueous cavities of significant extension normal to the membrane plane. Although much of the water on the sample will have evaporated during the time that the grid was at high vacuum and not yet frozen, we presume that the water present in these cavities is either tightly bound to the protein or trapped by a thin layer of dried glucose. To obtain a more quantitative sense of these features, we calibrated the observed densities using the known scattering densities of lipid and protein, and by assuming that the lipid region represents acyl chains in a bilayer of $30\text{ \AA}$ thickness. We then estimate the thickness of the low-density wells to be $\sim 25\text{ \AA}$, consistent with aqueous pathways extending through a substantial fraction of the protein.

Because $\sim 50\%$ of EriC's mass is predicted to lie outside the transmembrane domains, the protein probably extends beyond the acyl-chain region of the lipid bilayer. Therefore, neither aqueous cavity is likely to be long enough to traverse the entire protein, but

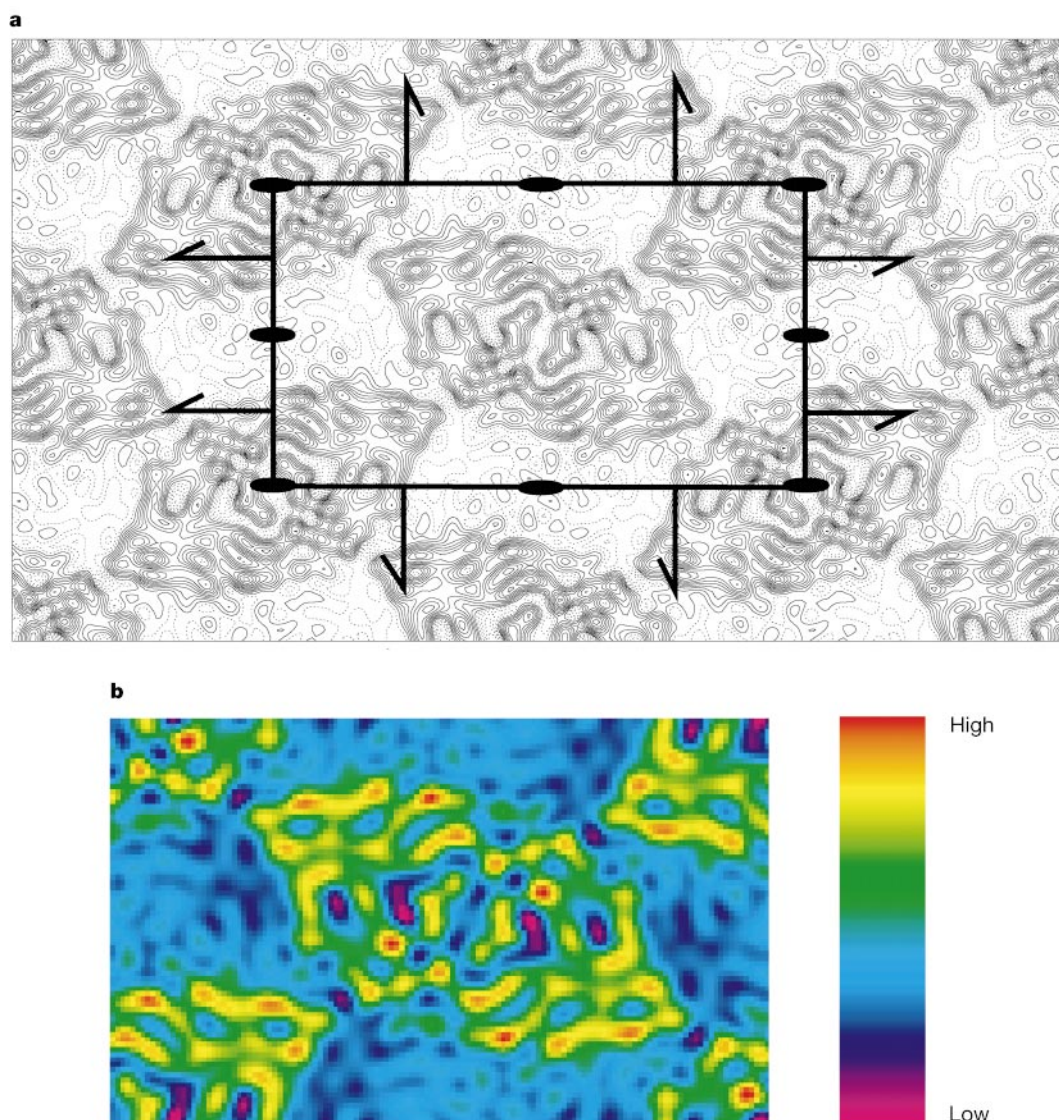


Figure 4 Projection density map of EriC at $6.5\text{ \AA}$ resolution. **a**, Contour map of EriC calculated from merged data obtained from seven crystals. A unit cell is boxed with the a axis ($140\text{ \AA}$) horizontal and the b axis ($80\text{ \AA}$) vertical in two-sided plane group $P2_22_1$. The two-fold axes perpendicular to the membrane plane and the screw axes parallel to the membrane plane are indicated (with the central two-fold axis omitted for clarity). The zero

contour level is roughly the average density of the lipid area; negative and positive contours are shown as dotted and solid lines, respectively. **b**, Colour-coded EriC density map. A single unit cell is plotted using the same data as in **a** with densities colour coded according to the scale on the right.

the two taken together could comfortably span its width. For this reason, we propose that each pair of density wells represents a single pore kinked roughly halfway along its length, such that the intracellular and extracellular openings are offset when viewed in projection normal to the membrane plane. This model, which implies two such pores in the dimeric channel, provides a natural structural rationale for the double-barrelled gating behaviour of single CIC channels from eukaryotes, and supports the previous inference that CIC channel pores cannot be axially symmetric^{1,7}. In addition, with a separation of 25–40 Å, the two pores would, as expected, be far enough apart for them to be electrostatically independent at the physiological 10 Å Debye length¹.

The fact that EriC, a prokaryotic protein, provides a faithful structural explanation of the gating of eukaryotic CIC channels supports the proposal that double-barrelled architecture is a fundamental, evolutionarily conserved feature of the entire CIC family¹¹. The absence of a strong low-density well on the dimer's two-fold axis is inconsistent with a symmetric one-pore construction for this channel. Of course, a definitive picture of the pore's transmembrane path must await three-dimensional information emerging from images of tilted crystals.

Transmembrane topology of CIC channels is uncertain; hydrophobicity analysis predicts 10–12 membrane-spanning segments¹³, and the cytoplasmic location of amino and carboxy termini demands an even number of membrane crossings. At 6.5 Å resolution, untilted transmembrane α -helices should produce strong circular density peaks in projection, and four such peaks are easily discerned in each EriC monomer. Additional elongated density peaks may represent tilted helices or other secondary structure. We count 12 peaks in all, but this projection map is insufficiently detailed to distinguish 10 from 12, and three-dimensional information will be required to identify all the transmembrane helices. Previous work on mammalian CIC-1 (ref. 9) characterized an externally exposed, putative pore-lining residue just preceding the fourth transmembrane segment (D4) that can be disulphide-cross-linked to its dimeric twin when substituted by cysteine, a reaction that localizes this residue near the two-fold axis of the dimer. In the EriC structure only one prominent density peak lies close to the two-fold axis in each monomer; notably, this density also abuts the larger of the pore openings. We thus speculate that this strong density represents D4. If this region undergoes a conformational change with gating, residues that line the pore in the open state might face the symmetry axis when the channel is closed. Therefore, crosslinking across the two-fold axis would lead to functional inhibition, as observed in CIC-1, by trapping a closed conformation of the channel. This conjecture predicts that the crosslinking should be dependent on conformational state, with reaction occurring only in the closed form of the channel. □

Methods

Crystal formation

EriCΔa, a decahistidine-tagged variant also lacking the five carboxy-terminal amino acids, was expressed in *Escherichia coli* through vector pASK-IBA2 (kindly provided by A. Skerra) using a tetracycline promoter, and was purified as described² with minor modifications. Pure EriC, with the His-tag proteolytically removed, was mixed with synthetic palmitoyl-oleyl-phosphatidylcholine (POPC) at a lipid/protein mass ratio of 0.4 (molar ratio ~50) in 4 mM decylmaltoide (DM). The mixture was dialysed against several changes of a solution containing 25 mM NaCl, 20 mM MgCl₂, 20 mM Tris-HCl, 0.8 mM NaN₃, pH 7.0. Crystals were collected and stored at 4°C. Samples were screened for crystallization by electron microscopy using 1% uranyl acetate as negative stain. Unstained specimens were prepared for high-resolution microscopy by deposition on carbon-coated copper grids, followed by embedding in 1% glucose with 0.25 mg ml⁻¹ bacitracin as a wetting agent. Samples were immediately transferred to the electron microscope and cooled by liquid nitrogen *in situ* in a Gatan cryo-stage.

³⁶Cl⁻/³H[glutamate] double-label flux assay

The two-dimensional crystals were collected by centrifugation at 3,000g (5 min), washed with dialysis buffer, and then extracted with 10 mM DM (room temperature, 1 h). Equilibrium exchange of vesicle-entrapped ³⁶Cl⁻ was measured as described². Briefly, EriC was reconstituted at varying concentrations into vesicles (500 µg lipid per sample) in the presence of the radioactive tracers ³⁶Cl⁻ and ³H[glutamate]. ³⁶Cl⁻ release was initiated by diluting the vesicles ninefold into tracer-free solution, and the entrapped ³⁶Cl⁻/³H[glutamate] counts were determined 2.5–3.5 h after dilution.

Electron microscopy

Electron micrographs were recorded in the Hitachi HF2000 FEG microscope at the MRC, Cambridge, England, at an accelerating voltage of 200 kV and nominal magnification of ×60,000. Magnification was later calibrated with tobacco mosaic virus. Unstained, glucose-embedded samples were imaged under low-dose conditions, ~5–15 electrons per Å², using Kodak SO163 film. Micrographs were screened by optical diffraction, and only those with evenly distributed spots extending to at least 10 Å resolution were used for final analysis.

Image processing

Well-ordered areas of 5,000 × 5,000 pixels were scanned at 7 µm per pixel using a Zeiss SCAI scanner, and computationally reduced to 14 µm per pixel. Data were corrected for lattice distortions, the contrast transfer function (CTF) and astigmatism using the MRC image processing programs^{14,15}. Two passes of unbending were used, with a reference area of 200 × 200 pixels in the first pass and 150 × 150 pixels in the second. The image amplitudes were scaled as a function of resolution to compensate for resolution-dependent attenuation, using the program SCALIMAMP3D with bacteriorhodopsin as reference¹⁶.

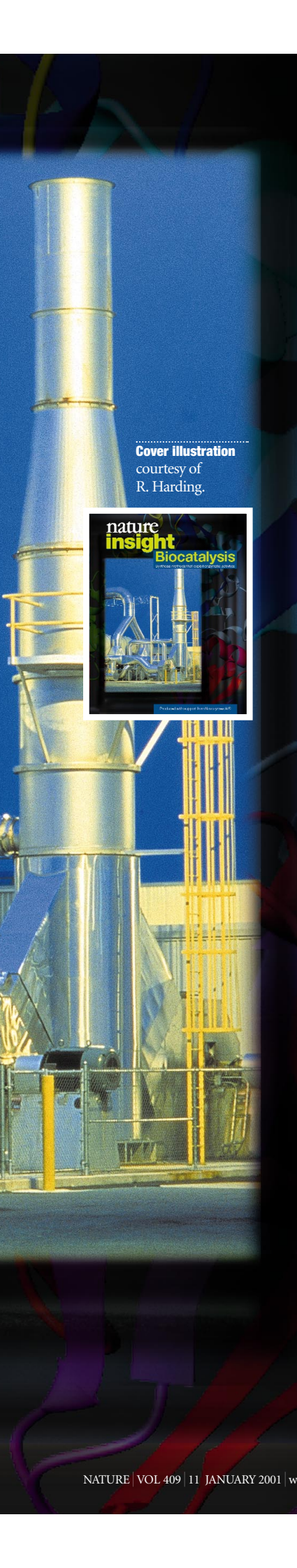
Received 25 August; accepted 30 October 2000.

- Maduke, M., Miller, C. & Mindell, J. A. A decade of CLC chloride channels: structure, mechanism, and many unsettled questions. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 411–438 (2000).
- Maduke, M., Pheasant, D. J. & Miller, C. High-level expression, functional reconstitution, and quaternary structure of a prokaryotic CIC-type chloride channel. *J. Gen. Physiol.* **114**, 713–722 (1999).
- Jentsch, T. J. & Gunther, W. Chloride channels: an emerging molecular picture. *BioEssays* **19**, 117–126 (1997).
- Ludewig, U., Pusch, M. & Jentsch, T. J. Two physically distinct pores in the dimeric CIC-0 chloride channel. *Nature* **383**, 340–343 (1996).
- Ludewig, U., Pusch, M. & Jentsch, T. J. Independent gating of single pores in CLC-0 chloride channels. *Biophys. J.* **73**, 789–797 (1997).
- Miller, C. Open-state substructure of single chloride channels from Torpedo electroplax. *Phil. Trans. R. Soc. Lond. B* **299**, 401–411 (1982).
- Middleton, R. E., Pheasant, D. J. & Miller, C. Homodimeric architecture of a CIC-type chloride ion channel. *Nature* **383**, 337–340 (1996).
- Saviane, C., Conti, F. & Pusch, M. The muscle chloride channel CIC-1 has a double-barrelled appearance that is differentially affected in dominant and recessive myotonia. *J. Gen. Physiol.* **113**, 457–468 (1999).
- Fahlke, C., Rhodes, T. H., Desai, R. R. & George, A. L. Jr. Pore stoichiometry of a voltage-gated chloride channel. *Nature* **394**, 687–690 (1998).
- Valpuesta, J. M., Carrascosa, J. L. & Henderson, R. Analysis of electron microscope images and electron diffraction patterns of thin crystals of phi 29 connectors in ice. *J. Mol. Biol.* **240**, 281–287 (1994).
- Middleton, R. E., Pheasant, D. J. & Miller, C. Purification, reconstitution, and subunit composition of a voltage-gated chloride channel from Torpedo electroplax. *Biochemistry* **33**, 13189–13198 (1994).
- Grigorieff, N., Beckmann, E. & Zemlin, F. Lipid location in deoxycholate-treated purple membrane at 2.6 Å. *J. Mol. Biol.* **254**, 404–415 (1995).
- Jentsch, T. J., Steinmeyer, K. & Schwarz, G. Primary structure of Torpedo marmorata chloride channel isolated by expression cloning in *Xenopus* oocytes. *Nature* **348**, 510–514 (1990).
- Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. & Zemlin, F. Structure of purple membrane from halobacterium halobium: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. *Ultramicroscopy* **19**, 147–178 (1986).
- Crowther, R. A., Henderson, R. & Smith, J. M. MRC image processing programs. *J. Struct. Biol.* **116**, 9–16 (1996).
- Schertler, G. F., Villa, C. & Henderson, R. Projection structure of rhodopsin. *Nature* **362**, 770–772 (1993).

Acknowledgements

We are grateful to R. Henderson for making facilities at the MRC, including the Hitachi microscope, available to us for low-dose experiments, to J. Berriman for help and hand-holding, and to T. Walz for crucial advice on crystallization. We also acknowledge D. Pheasant and L. Melanson for technical assistance. Support was provided in part by a grant to J.A.M. from the NIH.

Correspondence and requests for materials should be addressed to J.A.M. (e-mail: jmindell@brandeis.edu).

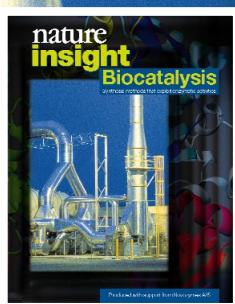


nature insight

Biocatalysis

Synthesis methods that exploit enzymatic activities

Cover illustration
courtesy of
R. Harding.



Biocatalysis underpins some of the oldest chemical transformations known to humans, for brewing predates recorded history. The Sumerians, for instance, produced at least 19 different types of beer. This practical art was the fuse for the explosion in understanding of organic and biological chemistry that took place in the nineteenth century. Coining the word ‘catalysis’, Berzelius divined that it must play a central role in life’s processes: “in the living plants and animals thousands of catalytic processes go on between the tissues and the fluids, and produce the amount of dissimilar chemical syntheses for whose formation from the common raw material...we could never see acceptable cause.”

Studies of fermentation led to key insights into life’s chemistry by Liebig, Pasteur and Emil Fischer, among others, culminating in the identification of enzymes (‘in yeast’) as nature’s catalytic molecules and Fischer’s intuitive leap of the ‘lock and key’ mechanism for their specificity.

It is this specificity that draws the interest of chemists seeking selective catalytic agents. But the trials of putting biocatalysis to industrial use are amply illustrated by the attempts in 1941 to produce fungal penicillin in what was basically a whole-cell process. It yielded such small amounts that the antibiotic had to be collected and recycled from the first patient’s urine.

This Insight shows just how far things have progressed since then. The diversity of potentially useful enzymes at the chemist’s disposal is now vast, supplemented by catalytic RNAs and antibodies. On page 226 Walsh surveys this arsenal, and discusses its deployment in applications ranging from chiral resolution to bioremediation of pollution. Koeller and Wong describe on page 232 how enzymes can become practical tools for the organic chemist, offering solutions to synthetic problems that seem intractable to artificial catalysts. Traditionally, enzymes have been regarded as catalysts designed to work in water. But on page 241 Klivanov shows how some can develop altered selectivities and enhanced thermal stability in nonaqueous solvents. On page 247 Khosla and Harbury explain how modular enzymes can be reshuffled or augmented to develop new functions in a rational manner. But ‘rationality’ is not the only answer to enzyme design, and Arnold shows on page 253 that *in vitro* evolution techniques provide the means to ‘breed’ and optimize new, non-natural enzymes. Whether or not a particular enzyme will deliver on its industrial potential depends, however, on a host of factors. On page 258 Witholt *et al.* provide an industry-wide perspective on the current successes and future challenges of using biocatalysts on a commercial scale.

We are pleased to acknowledge the financial support of Novozymes A/S in producing this Insight. As always, though, *Nature* carries sole responsibility for all editorial content and peer-review. We hope that readers will find this collection of reviews informative and thought provoking.

Philip Ball Consultant Editor

Karl Ziemelis Physical Sciences Editor

Liz Allen Publisher

Enabling the chemistry of life

Christopher Walsh

Biological Chemistry and Molecular Pharmacology Department, Harvard Medical School, Boston, Massachusetts 02115, USA

Enzymes are the subset of proteins that catalyse the chemistry of life, transforming both macromolecular substrates and small molecules. The precise three-dimensional architecture of enzymes permits almost unerring selectivity in physical and chemical steps to impose remarkable rate accelerations and specificity in product-determining reactions. Many enzymes are members of families that carry out related chemical transformations and offer opportunities for directed *in vitro* evolution, to tailor catalytic properties to particular functions.

The myriad chemical transformations carried out by every living organism are enabled by hundreds to thousands of proteins (enzymes) and, less frequently, RNAs (ribozymes), which have catalytic activity for conversion of a particular set of substrates to specific products. Some of these reactions are carried out by related families of protein biocatalysts, which act generically in the same way but exert specific recognition for transformation of a particular substrate molecule. For example, the orderly control of the location and lifetime of proteins in cells is managed by dozens of related proteases that hydrolyse peptide bonds of protein substrates in ways that are controlled in time and space. Proteases can be exquisitely specific for a particular peptide bond in a protein substrate, or they can be relentlessly nonspecific: the former set of proteases are involved in turning on biological signals, the latter in the clean-up phases of degradation and protein turnover.

When cells respond to external messenger molecules, such as the protein growth factors and hormones erythropoietin and insulin, or small-molecule hormones such as adrenaline or prostaglandins, signalling pathways are set in motion by catalytic action of cascades of protein kinases. The protein kinases are built from a small set of architectural types, and all catalyse phosphoryl transfer from ATP to the side-chain hydroxyl of serine, threonine or tyrosine residues. There are hundreds of such kinases in animal genomes. Selectivity is imposed on this generic chemical phosphorylation reaction by protein–protein interactions between a given kinase and its protein substrate and by cascades of such kinase/protein substrate pairs that ultimately lead to changes in activity and location of proteins, and to selective gene activation.

In addition to the large number of enzymes that act on macromolecular protein substrates, there are also enzymes that engage in truly sophisticated chemistry on small

Figure 1 Diverse chemical reactions facilitated by biocatalysts. ACC lyase, 1-aminocyclopropane-1-carboxylate lyase; IPNS, isopenicillin N synthase; OMP decarboxylase, orotidine-5'-phosphate decarboxylase.

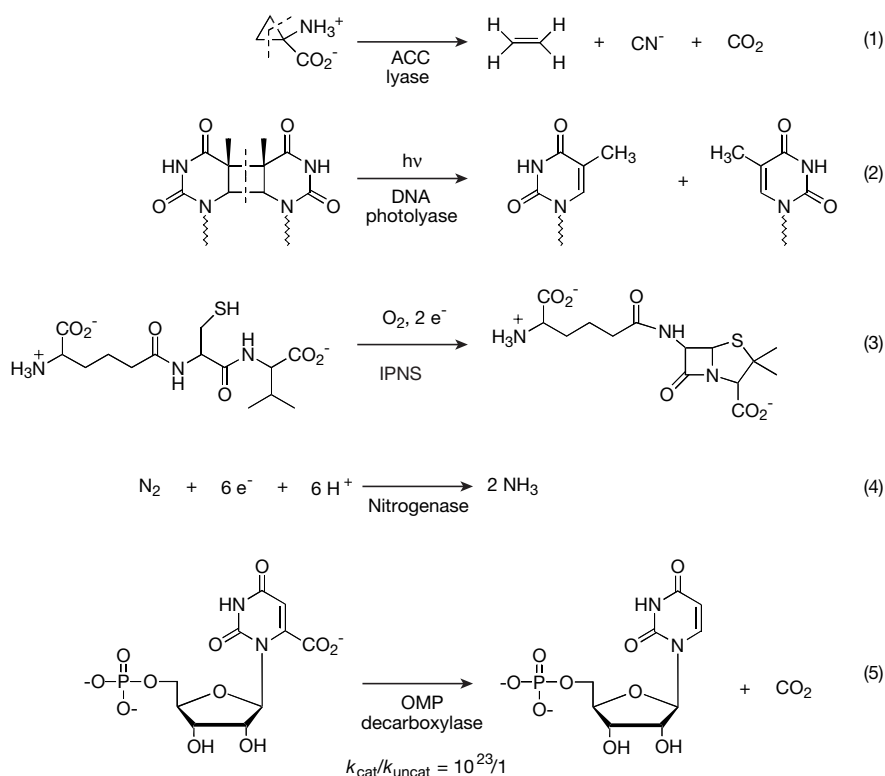
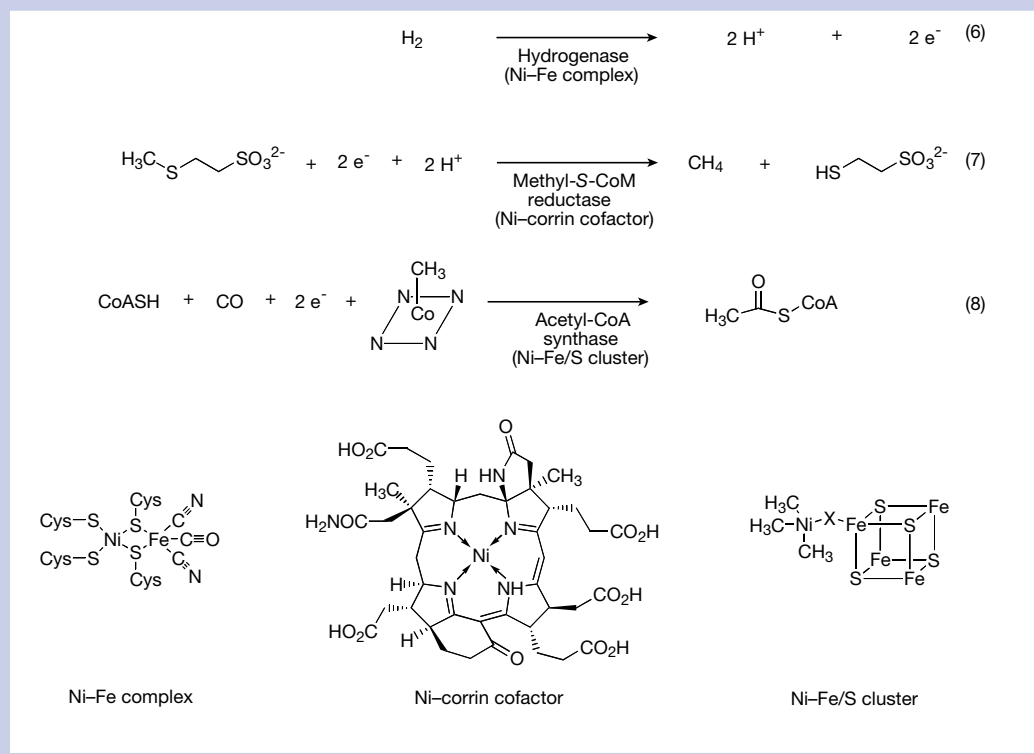


Figure 2 Nickel-based enzymatic transformations in methanogenic archaeobacteria.



organic molecules. The fragmentation of 1-aminocyclopropane-1-carboxylate to the fruit-ripening hormone ethylene¹, the photon-induced 2 + 2 cycloreversion of thymine dimers to repair DNA damaged by ultraviolet light², the bis-cyclization of the tripeptide aminoadipoyl-cysteinyl-D-valine (ACV) to isopenicillin N (ref. 3), and the reduction of dinitrogen (N₂) to two molecules of ammonia (NH₃) during nitrogen fixation⁴ are just a few examples of the range of biological chemistry facilitated by biocatalysts (Fig. 1). Enzymes as biocatalysts are remarkable not only in themselves, but also for the inspiration and guidance they provide to synthetic organic and inorganic chemists striving to reproduce and expand nature's chemical repertoire. Several of the useful attributes of biocatalysts, such as their use as reagents for chemical synthesis and scale-up, and directed evolution to tailor chemical transformations, are explored in other articles in this Insight.

Biocatalysts and their *ex vivo* utility

Biocatalysts carry out the chemistry of life, the controlled chemical transformations in primary metabolism and the generation of natural-product diversity in secondary metabolism of plants and microbes. Classically, the subset of proteins with catalytic activity — the enzymes — has been the focus of biocatalysis research. But there is an increasing focus on catalytic RNA (ribozymes), the discovery of which in the 1980s supported the arguments for an 'RNA world'^{5,6} antecedent to the contemporary world where proteins are the workhorse biocatalysts. Most recently, Joyce and co-workers⁷ have reported catalytic DNA molecules, and directed evolution of both RNA and DNA biocatalysts will continue to expand their potential. The current set of RNA and DNA catalysts have been assayed and developed for activities in nucleic-acid replication and in protein synthesis^{8,9}, but it remains to be seen how suitable they will be for the chemically diverse reactions encompassed by existing enzyme catalysts.

The twin hallmarks of enzyme biocatalysts are the remarkable specificities and sometimes phenomenal rate accelerations achieved. A typical enzyme, with a relative molecular mass of 50,000 (*M_r* 50K), is comprised of 450 amino-acid residues: 19 chiral L-amino acids and glycine. If glycine makes up 10% of the residues, then there are at least 400 residues with chiral centres to provide an asymmetric microenvironment for substrate binding and subsequent chemical

transformation in the enzyme's active site. This is the underlying structural basis for the action of all enzymes as chemoselective and regio- and stereospecific catalysts. In terms of rate accelerations, the relative values over nonenzymatic rates of transformation are often 10¹⁰, for example for protease-mediated hydrolysis of peptide bonds, and can reach 10²³ in the example of orotidine decarboxylase in the pyrimidine biosynthetic pathway¹⁰ (reaction 5 in Fig. 1). In absolute terms, enzymes have turnover numbers from as slow as one catalytic event per minute to 10⁵ per second (as in the hydration of CO₂ to HCO₃⁻ by carbonic anhydrase)¹¹.

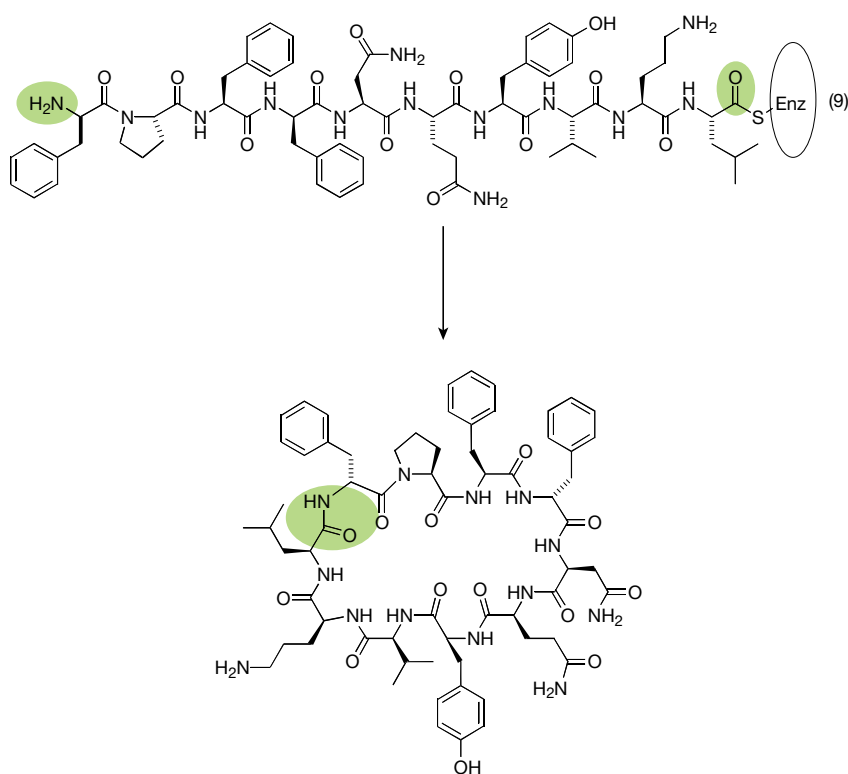
These two attributes of enzymatic biocatalysts have spurred much investigation into both the structural and mechanistic bases of the chemical transformations and have stimulated much of the study of enzymes in chemical synthesis (see review in this issue by Koeller and Wong, pages 232–240). *In vivo*, enzymes operate in buffered aqueous environments with ionic strength and pH control, although microbes that live at extremes of temperature and pH are of particular current interest because of the stability of their constituent enzymes. Much attention in biocatalyst process design (see accompanying review by Witholt *et al.*, pages 258–268) is on how to prolong useful lifetimes of enzyme catalysts and to have them operate in media not ordinarily compatible with life.

The past two decades have also witnessed an intense exploration of catalytic antibodies¹². To prepare these antibodies, ligands are synthesized that typically mimic transition states of particular chemical transformations, such as ester hydrolysis, amide synthetase and Claisen condensation. Monoclonal antibodies are then selected that display high-affinity binding to the ligands, thus enriching for antibody proteins with a binding-site geometry complementary to the shape of the true transition state. Some of the antibodies selected in this way show catalysis of the desired reactions, with the selectivity and rate accelerations expected for chiral protein-based catalysts^{13,14}. But low catalytic turnover numbers have so far limited the use of catalytic antibodies in chemical synthesis or process work.

Biocatalysts or biomimetic catalysts?

With their unerring stereoselectivity and high catalytic efficiency, nature's enzymatic catalysts have been a stimulus and counterpoint

Figure 3 Cyclization catalysed by the thioesterase domain of tyrocidine synthetase.



for generations of chemists who have designed and tested bioorganic and bioinorganic versions of biomimetic catalysts, whether for example to mimic macrocyclizations of natural products or to produce analogues of hydrogenase or nitrogenase catalysts or the photosynthetic splitting of water¹⁵. The mimics may operate under harsher solvent and temperature conditions, and may be more robust in terms of lifetime (if not throughput per catalyst molecule). When organic coenzymes (such as flavins, pyridoxal or thiamin) or inorganic cofactors (iron/sulphur clusters, metalloporphyrins) are crucial components of the enzymatic catalysis, the biomimetic and natural catalysts often show design convergence and may recapitulate some of the steps in biocatalyst evolution. The three nickel enzymes in methanogenic bacteria (thought to be contemporary descendants of primordial organisms), which carry out nickel-based hydrogenation, nickel-based methyl thioether reduction to methane, and nickel-based carbonylation of a methyl co-substrate to produce acetate, can be viewed as such an intersection^{16,17} (Fig. 2).

When is it worthwhile for the synthetic or process chemists to reject synthetic reagents and catalysts in favour of enzymes to carry out a specific transformation? This may vary with individual preference and each case must be judged on its own merits. Lipases and other hydrolases have clear advantages in kinetic resolutions of intermediates (see below), penicillin acylases have long been a mainstay of semisynthetic processes in the β -lactam antibiotic industry, and enzymatic aldol condensations have shown their worth in complex oligosaccharide syntheses¹⁸.

Chemical transformations well suited to enzymes

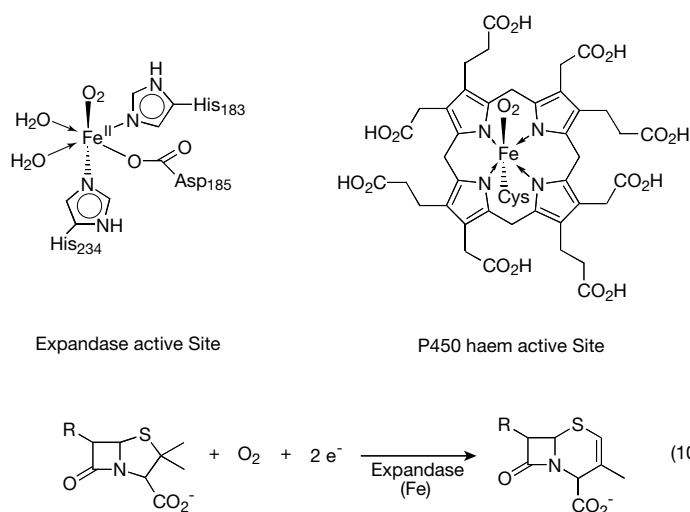
The accompanying review by Khosla and Harbury (pages 247–252) explores the multimodular enzymes that function as molecular solid-state assembly lines for the generation of thousands of polyketide natural products and non-ribosomal peptide antibiotics, including important medicinal compounds such as erythromycin, rapamycin, epothilone, lovastatin, penicillins, cyclosporin and vancomycin^{19–21}. These sequentially elongating acyl transfers seem

particularly apt loci for use as enzymatic rather than biomimetic catalysis. Some of the assembly lines, such as those for erythromycin or cyclosporin, produce the intramolecularly cyclized macrolactones or macrolactams. It has recently been shown²² that the last 30K (thioesterase) domain of the 724K protein assembly line of tyrocidine synthetase retains the ability to cyclize 9–11-residue peptidyl thioesters with regio- and stereoselectivity, raising the prospect for practical enzymatic macrocyclizations by a robust, small protein fragment (Fig. 3, reaction 9).

The reprogramming of the component enzyme domains of these assembly lines to create new, unnatural ‘natural’ products is one of the goals of combinatorial biosynthesis. The order of the enzymatic domains in the assembly lines specifies which monomer substrates are activated, condensed and elongated. So altering the order and permutations of these domains offers the chance to control product structure. The directed evolution of the catalytic domains of polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) assembly lines by gene shuffling and other approaches (see accompanying review by Arnold, pages 253–257) can create designed diversity in complex natural products.

Once the nascent products have been released from the PKS and NRPS assembly lines, the polyketide or polypeptide may require further enzymatic transformations to attain antibiotic properties. This is the case for penicillins, vancomycin and erythromycin, to cite just three important examples¹⁹. Baldwin and co-workers^{23,24} showed that the tripeptide ACV is oxidatively transformed to the 4–5 bicyclic β -lactam ring system by isopenicillin *N* synthase (IPNS; Fig. 1, reaction 3). IPNS is a member of a substantial family of iron-containing enzymes that use Fe^{2+} to activate both O_2 and the specific co-substrate for complex redox chemistry²⁵. In IPNS, both atoms of dioxygen are reduced to water and the ACV tripeptide undergoes four-electron oxidation and directed C–S bond and C–C bond formation as the β -lactam forms. A cousin of IPNS, the expandase enzyme, is used by cephalosporin-producing organisms to expand the five-membered ring in penicillins to the six-membered ring in cephalosporin

Figure 4 Comparison of expandase active site with a typical haemprotein oxygenase.



antibiotics (Fig. 4, reaction 10). The ligand set around the active-site iron — one Glu, two His residues — is the same, but the reaction flux is distinct (Fig. 4). Other members of this non-haem dioxygenase family include the enzyme responsible for hydroxylating prolyl residues in procollagen to predispose it to triple-helix formation in mature collagen, the most abundant protein in the human body. There are clear potential benefits to understanding the molecular basis for how the high-valent oxo-iron reagents are controlled and directed to flawlessly different chemical outcomes in the members of this redox enzyme family, so that they might be subjected to *in vitro* evolution to generate new reaction fluxes.

Many natural products, from morphine and codeine to vancomycin, undergo oxidative cyclization reactions that are regio- and stereospecific and seem to be mediated by a different superfamily of iron-containing oxidases, the cytochromes P450, with Fe^{2+} embedded in a haem macrocycle (Fig. 4). Protein superfamilies are groups of proteins with distinct chemical functions, amino-acid sequences of recognizable but sometimes marginal homology, and convergent three-dimensional structures. In the vancomycin family of glycopeptide antibiotics there are three crosslinks that convert an acyclic heptapeptide, the product of the NRPS assembly line, into a rigid scaffold, crosslinked at Tyr₂-PheGly₄-Tyr₆ and PheGly₅-dihydroxyPheGly₇ (Fig. 5, reaction 11). There are three P450 cytochromes in the biosynthetic gene cluster; each might enact a regiospecific phenolic crosslink. Harnessing such catalysts for related transformations might lead to new vancomycins.

Several natural products contain tandem five-membered-ring heterocycles (oxazoles and thiazoles) that arise from enzymatic cyclization of serine or cysteine residues in peptide precursors²⁶. These include the *Escherichia coli* antibiotic microcin B17, which kills neighbouring bacteria by poisoning the enzyme DNA gyrase and thus blocking DNA replication, in much the same way as does the best-selling antibiotic ciprofloxacin²⁷ (Fig. 6). Such heterocycles are also found in the iron-chelating siderophores that act as virulence factors in infections by *Pseudomonas aeruginosa*, *Vibrio cholerae* and the causative agent of the black plague, *Yersinia pestis*^{28,29}. Enzymes that heterocyclize serine, threonine and cysteine side chains in peptides (Fig. 6, reactions 12, 13) may create either DNA-seeking or iron-chelating sites in any peptide library that could then be screened for biological activity.

Superfamilies, genomics and enzyme evolution

The iron-containing dioxygenases that include IPNS and expandase, and the cytochrome P450 variants that introduce crosslinks, comprise redox enzyme superfamilies that are good candidates for

engineering for altered catalytic properties and specificities. Genomic and proteomic searches can identify many enzyme superfamily members through amino-acid sequence homologies, in which scaffolding and structural architecture will be predictable. Some of these proteins are of unknown ('orphan') function, and the assignment of function is one of the major postgenomic challenges of proteomic research. Recent cases in the crotonase superfamily (Fig. 7, reactions 14–16) and enolase superfamily (Fig. 7, reactions 17–19)^{30–32} indicate that the active sites all generate carbanionic transition states from bound substrates and then use carbanion chemistry for directed fluxes and distinct product outcomes. These families should be fruitful starting points for directed enzyme evolution to elicit new fluxes, based on the knowledge that carbanion chemistry will be facilitated

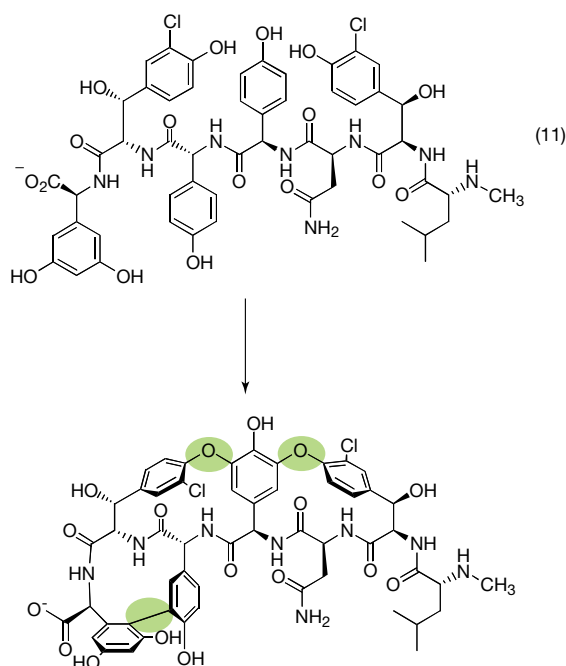
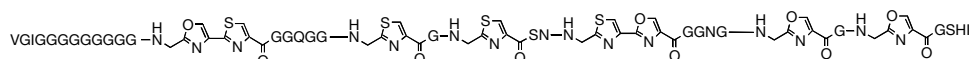


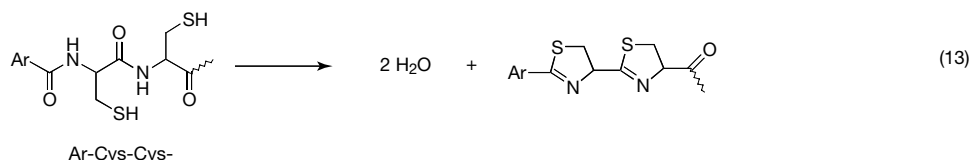
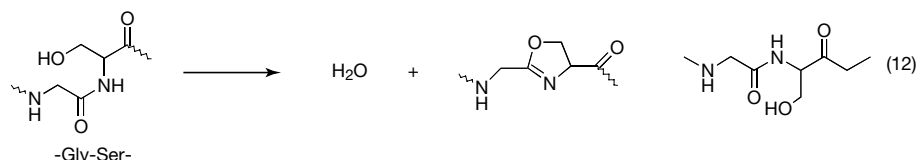
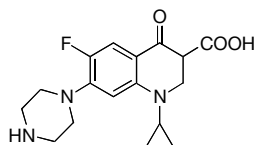
Figure 5 Crosslinking by cytochrome P450 enzymes to produce the vancomycin Aglycone.

Figure 6 DNA gyrase inhibitors and biosynthesis of peptide heterocycles.

MccB17



Ciprofloxacin



in one of the co-substrates and that binding sites can be re-engineered for electrophilic substrate components.

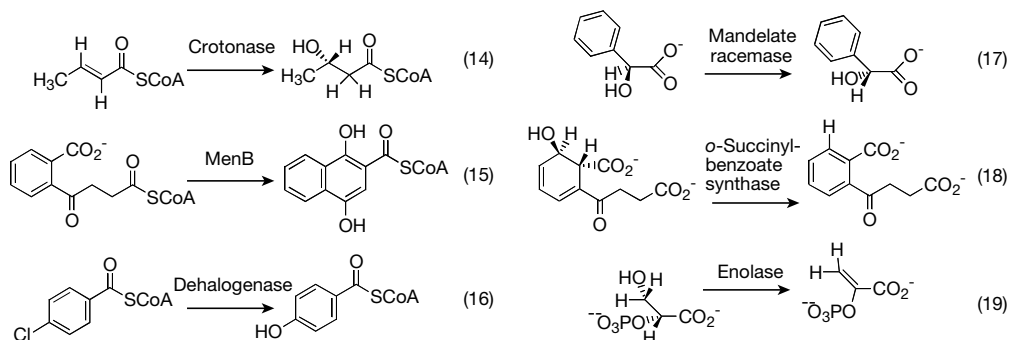
Once an enzyme has been evolved to have a detectable and desirable new activity³³, additional rounds of *in vitro* evolution can improve its stability and robustness. The biological selection methods are sufficiently powerful that one can find outcomes that are very rare biologically in a short space of time. A good example is the recent report³⁴ of expression of a functional carotenoid biosynthetic pathway in *E. coli* by selecting for bacteria that become red. The continuing progress in biological production of polyhydroxyalkanoate polymers with controlled sizes and properties³⁵ by engineering the respective polymerases increases the likelihood of economically viable production of these biodegradable plastics by biocatalysis.

Enzymes in bioremediation

One of the most active areas of applied enzymology in the past two decades has been the study of enzymes capable of bioremediation: the breakdown of organic and inorganic pollutants. There are now substantial databases of enzymes and the bioremedial transforma-

tions³⁶ they catalyse, which include the breakdown of aromatic and heteroaromatic pollutants by oxidative, reductive and hydrolytic transformations. Iron-containing dioxygenases and monooxygenases, with overlapping regio- and chemospecificities, are superfamilies that represent good starting points for application of many of the strategies noted here and in the specific accompanying articles for directed enzyme evolution to broaden substrate recognition. It is likely that bioremediation scenarios in the field will require the tailored enzymes to work in their host microbial cells rather than as *ex vivo* catalysts. Engineering of multistep metabolic pathways by introducing heterologous genes³⁷ and *in vivo* expression may well be required for efficient degradation of non-biogenic compounds. As many waste sites have a witches' brew of foreign compounds, multiple pathways engineered stably into a microbe or, more probably, mixed bacterial communities that can coexist stably, will be required. The enzymology of processing of toxic inorganic ions has also progressed in recent years to include mercury, copper, cadmium, silver, arsenic and cobalt. This might ultimately make remediation schemes for inorganic pollutants feasible³⁸.

Figure 7 Representative reactions catalysed by the crotonase superfamily and the enolase superfamily.



Conclusions

As structural genomics continues to reveal the folds and scaffolds of several members of all the principal superfamilies of enzymes, the molecular bases of recognition of substrates and directed fluxes through specific transition states to particular subsets of products will become increasingly clarified. In turn this will aid in enzyme evolution to select and detect new activities and then to incorporate improved catalytic efficiency, attributes of specificity, and structural features optimized to a given operating microenvironment. For both *in vitro* applications for a specific synthetic chemical step and for *in vivo* construction of new metabolic pathways, the applications for enzymes in practical biocatalysis will continue to burgeon. Small-molecule chemical transformation catalysed by enzymes from microorganisms that live in unusual environments or conduct chemical warfare against their neighbours have been and are likely to remain good hunting grounds for new enzyme transformations. □

- Peiser, G. *et al.* Formation of cyanide from carbon 1 of 1-aminocyclopropane-1-carboxylic acid during its conversion to ethylene. *Proc. Natl Acad. Sci. USA* **81**, 3059–3063 (1984).
- Sancar, A. Structure and function of DNA photolyase. *Biochemistry* **33**, 2–9 (1994).
- Schofield, C. J. *et al.* Proteins of the penicillin biosynthesis pathway. *Curr. Opin. Struct. Biol.* **7**, 857–864 (1997).
- Bertino, I., Gray, H. B., Lippard, S. J. & Valentine, J. S. *Bioinorganic Chemistry* (University Science Books, Mill Valley, CA, 1994).
- Gesteland, R., Atkins, J. & Cech, T. R. (eds) *The RNA World* 2nd edn (Cold Spring Harbor Laboratory Press, 1999).
- Narlikar, G. J. & Herschlag, D. Mechanistic aspects of enzymatic catalysis: lessons from comparison of RNA and protein enzymes. *Annu. Rev. Biochem.* **66**, 19–59 (1997).
- Sheppard, T. L., Ordoukhanian, P. & Joyce, G. F. A DNA enzyme with N-glycosylase activity. *Proc. Natl Acad. Sci.* **97**, 7802–7807 (2000).
- Zhang, B. & Cech, T. R. Peptide bond formation by *in vitro* selected ribozyme. *Nature* **390**, 96–100 (1997).
- Cech, T. R. & Golden, B. L. in *The RNA World* 2nd edn (eds Gesteland, R., Atkins, J. & Cech, T. R.) 321–349 (Cold Spring Harbor Laboratory Press, 1999).
- Radzicka, A. & Wolfenden, R. A proficient enzyme. *Science* **267**, 90–93 (1995).
- Stryer, L. *Biochemistry* 4th edn (Freeman, San Francisco, 1995).
- Patten, P. A. *et al.* The immunological evolution of catalysis. *Science* **271**, 1086–1091 (1996).
- Wagner, J. A., Lernner, R. A. & Barbas, C. F. III Efficient aldolase catalytic antibodies that use the enamine mechanism of natural enzymes. *Science* **270**, 1797–1800 (1995).
- Smithrud, D. B. & Benkovic, S. J. The state of antibody catalysis. *Curr. Opin. Biotechnol.* **8**, 459–466 (1997).
- Lippard, S. J. & Berg, J. M. *Principles of Bioinorganic Chemistry* (University Science Books, Mill Valley, CA, 1994).
- Walsh, C. T. & Orme-Johnson, W. H. Nickel enzymes. *Biochemistry* **26**, 4901–4906 (1987).
- Watt, R. K. & Ludden, P. W. Nickel binding proteins. *Cell Mol. Life Sci.* **56**, 604–625 (1999).
- Wong, C.-H. & Whitesides, G. M. *Enzymes in Synthetic Organic Chemistry* (Pergamon, Oxford, 1994).
- Cane, D. (ed.) Thematic issue on polyketide and nonribosomal peptide synthetases. *Chem. Rev.* **97**, 2463–2705 (1997).
- Konz, D. & Marahiel, M. How do peptide synthetases generate structural diversity? *Chem. Biol.* **6**, R34–R38 (1999).
- Cane, D. E., Walsh, C. T. & Khosla, C. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science* **282**, 63–68 (1998).
- Trauger, J., Kohli, R. M., Mootz, H., Marahiel, M. & Walsh, C. Peptide cyclization catalysed by the thioesterase domain of tyrocidine synthetase. *Nature* **407**, 215–218 (2000).
- Roach, P. L. *et al.* The crystal structure of isopenicillin N synthase, first of a new structural family of enzymes. *Nature* **375**, 700–704 (1995).
- Valegard, K. *et al.* Structure of a cephalosporin synthase. *Nature* **394**, 805–809 (1998).
- Que, L. One motif—many different reactions. *Nature Struct. Biol.* **7**, 182–184 (2000).
- SinhaRoy, R., Milne, J., Belshaw, P., Gehring, A. & Walsh, C. Oxazole and thiazole peptide biosynthesis. *Nat. Prod. Rep.* **16**, 249–263 (1999).
- Lewis, R. *et al.* Molecular mechanisms of drug inhibition of DNA gyrase. *BioEssays* **18**, 661–671 (1996).
- Quadrif, L. E., Keating, T. A., Patel, H. M. & Walsh, C. Assembly of the *Pseudomonas aeruginosa* nonribosomal peptide siderophore pyochelin: *in vitro* reconstitution of aryl-2,4-bis-thiazoline synthetase activity from PchD, E and F. *Biochemistry* **38**, 14941–14954 (1999).
- Gehring, A., Mori, L., Perry, R. & Walsh, C. The nonribosomal peptide synthetase HMWP2 forms a thiazoline ring during biogenesis of yersiniabactin, an iron-chelating virulence factor of *Yersinia pestis*. *Biochemistry* **37**, 11637–11650 (1998).
- Babbitt, P. C. *et al.* The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry* **35**, 16489–16501 (1996).
- Gerlt, J. A. & Babbitt, P. C. Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Curr. Opin. Chem. Biol.* **2**, 607–612 (1998).
- Hubbard, B. K. Functional and mechanistic investigations of enzymes in the enolase superfamily. Thesis, Univ. Illinois (2000).
- Tobin, M. B., Gustafsson, C. & Huisman, G. W. Directed evolution: the rational basis for irrational design. *Curr. Opin. Struct. Biol.* **10**, 421–427 (2000).
- Schmidt-Dannert, C., Umeno, D. & Arnold, F. Molecular breeding of carotenoid biosynthetic pathways. *Nature Biotechnol.* **18**, 750–753 (2000).
- Madison, L. L. & Huisman, G. J. Metabolic engineering of poly(3-hydroxyalkanoates): from DNA to plastic. *Microbiol. Mol. Biol. Rev.* **63**, 21–53 (1999).
- Wackett, L. P. *et al.* Predicting microbial degradation pathways. *Am. Soc. Microbiol. News* **65**, 87–94 (1999).
- McDaniel, R., Ebert-Khosla, S., Hopwood, D. A. & Khosla, C. Rational design of aromatic polyketide products by recombinant assembly of enzymatic subunits. *Nature* **375**, 549–554 (1995).
- Bizily, S. P., Rugh, C. L., Summers, A. O. & Meagher, R. B. Phytoremediation of methylmercury pollution: *merB* expression in *Arabidopsis thaliana* confers resistance to organomercurials. *Proc. Natl Acad. Sci. USA* **96**, 6808–6813 (1999).

Acknowledgements

Work cited from the author's laboratory has been supported by the National Institutes of Health. I thank B. Hubbard for drawing the artwork in figures 1–7.

Enzymes for chemical synthesis

Kathryn M. Koeller & Chi-Huey Wong

Department of Chemistry and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

New catalytic synthetic methods in organic chemistry that satisfy increasingly stringent environmental constraints are in great demand by the pharmaceutical and chemical industries. In addition, novel catalytic procedures are necessary to produce the emerging classes of organic compounds that are becoming the targets of molecular and biomedical research. Enzyme-catalysed chemical transformations are now widely recognized as practical alternatives to traditional (non-biological) organic synthesis, and as convenient solutions to certain intractable synthetic problems.

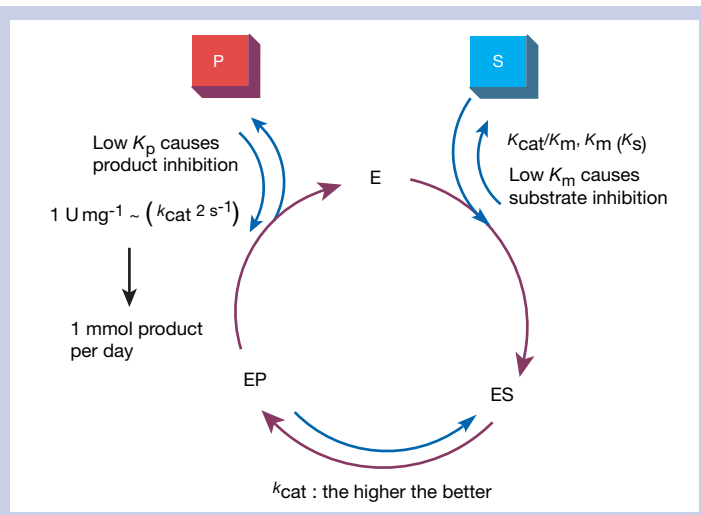
Most enzymes operate at room temperature, under neutral aqueous conditions, and in the absence of substrate functional-group protection. In organic synthesis, these biocatalysts can be used as the sole catalyst in a reaction, in combination with other enzymes, or with non-biological reagents. The chiral nature of enzymes results in the formation of stereo- and regiochemically defined reaction products with remarkable rate acceleration (typically 10^5 to 10^8). In addition, many enzymes accept unnatural substrates, and genetic engineering can further alter their stability, broaden their substrate specificity, and increase their specific activity. Molecules with several functional groups pose particular challenges to non-biological synthetic methods, but are natural targets for biological techniques. For example, large DNA and RNA molecules can be efficiently synthesized and manipulated by enzymatic processes, whereas equivalent chemical alternatives towards this end do not exist. Through the use of biocatalysts, otherwise impractical synthetic manipulations of complex molecules, such as carbohydrates, can be performed in an environmentally benign manner. Both natural and engineered enzymes can now be produced on a large scale in convenient host organisms using recombinant DNA technologies. The application of enzymes in synthesis thus represents a remarkable opportunity for the development of industrial chemical and pharmaceutical processes.

This article describes some of the recent developments in the rapidly growing field of enzymatic catalysis, with particular focus on the use of free enzymes (that is, outside the cell) in preparative asymmetric transformations. In many cases, free enzymes offer advantages over whole-cell processes, which may be more difficult to predict, control and manipulate. We shall give precedence to selected examples reported in the past five years; a broader overview can be found in other recent reviews^{1–15}.

Practical issues and limitations

Over 3,000 enzymes have so far been identified, and this number may be greatly augmented in the wake of genomic and proteomic research. The enzymes that have been exploited for organic synthesis, as well as the type of reaction catalysed, are summarized in Table 1. In general, several parameters affect the practicality of an enzymatic reaction (Fig. 1). Of particular importance are the specific activity (quantified by k_{cat}), specificity (determined by $k_{\text{cat}}/K_{\text{m}}$) and stability of the enzyme. In addition, the degree of inhibition by substrate or product (determined by their affinity to the enzyme) may be particularly important in the outcome of a reaction. In an ideal scenario, the enzyme used would have high specific activity and stability, and would be subject to minimal substrate and product inhibition. Furthermore, the extent of substrate specificity can determine whether a given enzyme will have general synthetic utility, with stereospecificity perhaps the most important parameter under consideration. Although enzymes with

Figure 1 Practical parameters to be considered in enzymatic synthesis. Substrate (S) and product (P) inhibitions occur when the corresponding dissociation constants (K_{s} and K_{p}) are too small (<0.1 mM). The enzyme specificity is determined by $k_{\text{cat}}/K_{\text{m}}$ (approximately the rate constant for substrate reacting with the enzyme) and the specific activity (k_{cat} , the rate constant from ES to EP) should be high enough for practical use. Typically an enzyme with $k_{\text{cat}} = 2 \text{ s}^{-1}$ will produce 1 mmol of product per day. For synthetic purposes, some reversible reactions have to be altered to irreversible reactions as they affect the enantiomeric purity of the product and the catalytic efficiency.



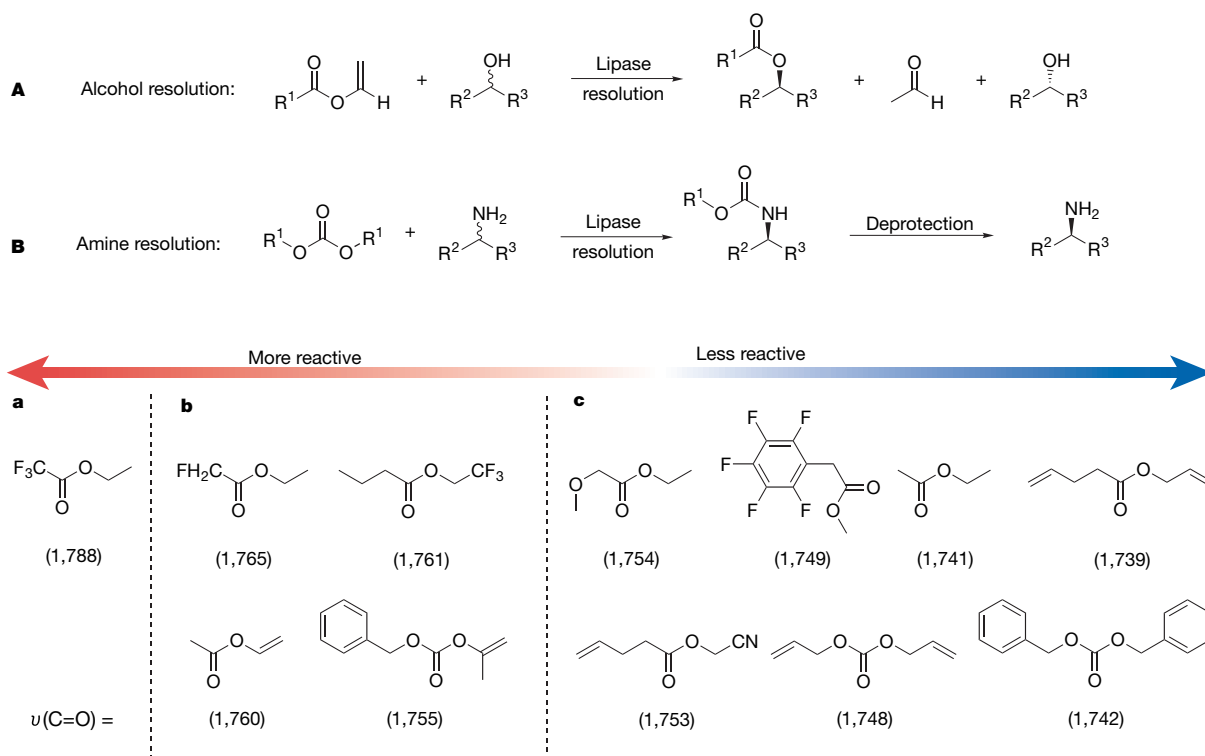


Figure 2 Enzyme-catalysed enantiotransformation. **A**, Lipase-catalysed resolution of alcohols using enol esters as irreversible transesterification reagents. **B**, Lipase-catalysed resolution of amines using temporary blocking groups. Reactivity of acylating reagents for amines, and their usefulness in enzymatic reactions are illustrated: **a**, too reactive for amines; **b**, useful for alcohols and good for amines under spontaneous reaction-suppressing conditions; and **c**, useful under reaction-promoting conditions. The number in parenthesis is infrared absorption maxima for the carbonyl group.

narrow substrate specificity are often efficient in catalysing reactions using their natural substrate, this property becomes a limitation when the development of catalysts for general purposes is the goal. Biocatalysts that accept a wide range of substrates to form enantiopure products are of particular interest to the synthetic chemist. Many enzymes have now proven synthetically useful and have become commercially available; however, one still cannot use enzymes for the formation of every desired linkage or resolution of any racemic mixture. Moreover, although many enzymes have been highly characterized with regard to substrate specificity and stereoselectivity, they may be unpredictable with unnatural substrates.

Hydrolytic enzymes in enantiotransformation

Obtaining enantiomerically pure intermediates and products efficiently and economically is of utmost importance in the pharmaceutical industry. Hydrolytic biocatalysts have been instrumental for these purposes. Esterases, lipases and proteases have been widely applied to the preparation of enantiopure compounds from racemic pairs, prochiral (precursors to chiral) or *meso* compounds, or diastereomeric mixtures^{1–13}. These enzymes are also active in organic solvents (see Klivanov's review in this issue, pages 241–246). The general premise behind enzymatic resolution is that the enzyme esterifies (or hydrolyses) only a single enantiomer of a racemic substrate, thus providing a means of separation. A particularly practical development in this respect is the use of enol esters as transesterification reagents¹⁵, which irreversibly force the enzymatic process in the forward direction (Fig. 2). This prevents loss of enantioselectivity resulting from the reverse reaction, and eliminates product-inhibition problems. Hydrolytic enzymes also effectively catalyse enantiocomplementary reverse hydrolysis (esterification, transesterification, aminolysis or amidation), providing access to both enantiomers of a desired

product^{2–4,6,13}. Equations and graphs developed by Sih *et al.*² for quantitative treatment of these enzymatic transformations allow prediction of enantiomeric excess^{2–4,7}. Further reaction optimization can be accomplished through molecular modelling of the active site, as developed by Kazlauskas¹⁶ and others. Modelling techniques can aid in prediction of the stereochemical course of the reaction, and also give insight into potential substrate or enzyme modifications that may increase selectivity.

One of the main advances in enzymatic resolution was realized through the use of hydrolytic enzymes in the presence of additional racemization catalysts^{17,18}. The drawback to the usual strategy of enzymatic resolution is that the desired enantiomer is obtained in a maximal 50% yield, as that is the composition of the enzyme's substrate in the racemic mixture. In dynamic resolution (reaction 1 in Box 1), on the

Table 1 Enzymes commonly used in organic synthesis

Enzymes	Reactions
Esterase, lipases	Ester hydrolysis, formation
Amidases (proteases, acylases)	Amide hydrolysis, formation
Dehydrogenases	Oxidoreduction of alcohols and ketones
Oxidases (mono- and dioxygenases)	Oxidation
Peroxidases	Oxidation, epoxidation, halohydrin
Kinases	Phosphorylation (ATP-dependent)
Aldolases, transketolases	Aldol reaction (C–C bond)
Glycosidases, glycosyltransferases	Glycosidic bond formation
Phosphorylases, phosphatases	Formation and hydrolysis of phosphate
Sulphotransferases	Formation of sulphate esters
Transaminases	Amino acid synthesis (C–N bond)
Hydrolases	Hydrolysis
Isomerases, lyases, hydratases	Isomerization, addition, elimination, replacement

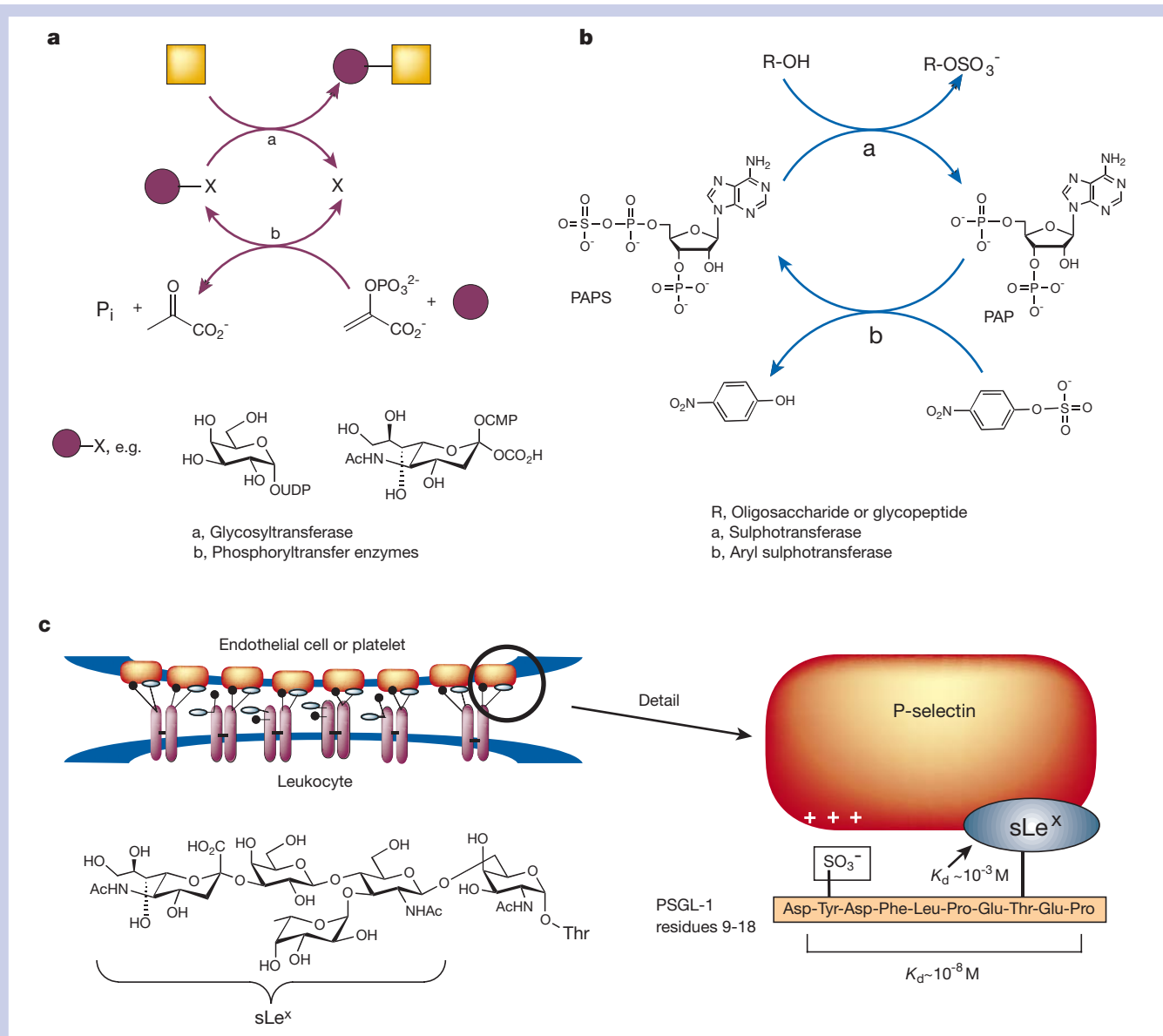


Figure 3 Cofactor regeneration. Regeneration of **a**, sugar nucleotides (monosaccharides denoted by circles and squares) and **b**, 3'-phosphoadenosyl-5'-phosphosulphate (PAPS) in enzymatic formation of glycosidic bonds and sulphate esters. **c**, Application to the synthesis of tyrosine sulphate-containing glycopeptides such as the N-terminal portion of P-selectin glycoprotein ligand-1 (PSGL-1).

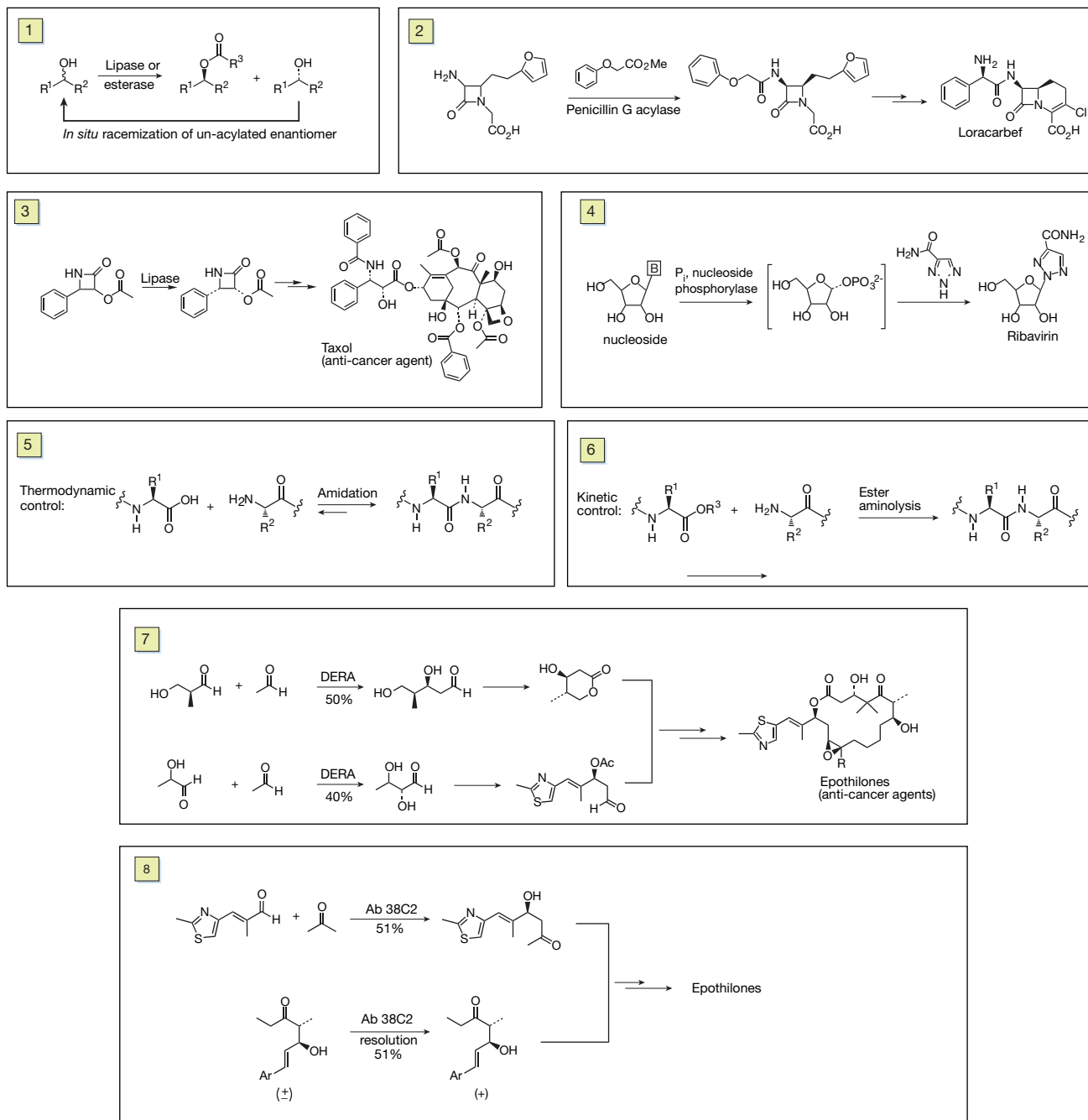
other hand, the enantiomer that does not serve as a lipase substrate is continually racemized to produce additional quantities of the enzymatic substrate. *In situ* racemization thereby gradually increases the overall concentration of the substrate recognized by the enzyme, resulting in product yield of greater than 50%. Notably, in the presence of ruthenium catalysts as racemizing agents, dynamic lipase resolutions have furnished products in high yield and >99% enantiomeric excess¹⁸.

Hydrolytic enzymes also provide efficient protecting-group strategies. Lipase-catalysed enantioselective reactions for temporary protection of amines have recently been reported¹⁹. The reactivity of a range of acylating reagents was observed to correlate with infrared carbonyl stretching frequency, resulting in the generation of a useful guideline for selection of alcohol- and amine-protecting groups (Fig. 2). Racemic amines first serve as substrates during lipase-catalysed resolution reactions, which produce enantiopure amines in protected form. As the acyl blocking groups have been chosen specifically to serve also as readily removable protecting groups, mild deprotection then yields chiral free amines. Furthermore, Waldmann and co-workers have pioneered the use of various esterases and lipases as

general reagents for deprotection under neutral conditions²⁰. Incorporation of these hydrolytic enzymatic reactions into synthetic schemes has allowed the preparation of various acid- and base-labile peptide conjugates that are not compatible with standard peptide protecting-group strategies. New phospho-, glyco- and lipopeptide conjugates have been assembled through the inclusion of enzymatic deprotection steps. The peptide conjugates thus produced have broad application in the study of signal transduction and controlled membrane localization²¹. Other practical syntheses using hydrolytic enzymes include: the use of penicillin acylase to cleave the side chain of β -lactam antibiotics and to introduce a new side chain, as illustrated by Eli Lilly's synthesis of an antibiotic (Box 1, reaction 2); the use of lipase in the Bristol-Myers Squibb synthesis of taxol (Box 1, reaction 3); and the use of nucleoside phosphorylases in Yamasa's synthesis of the antiviral Ribavirin (Box 1, reaction 4).

Cofactor-dependent enzymes and cofactor regeneration

A number of synthetically practical enzymatic reactions require cofactors, the costly nature of which precludes their use as stoichiometric reagents.

Box 1
Reactions

chiometric (non-regenerated) reagents. Instead, regeneration of the cofactor from its reaction by-product is necessary in order for the process to be economically and industrially feasible^{3,5}. Cofactor regeneration is also synthetically advantageous, as it drives the reaction to completion, prevents the accumulation of inhibitory cofactor by-products, simplifies the reaction work-up, and increases enantioselectivity. Several cofactors can be recycled effectively³, including nucleoside triphosphates such as ATP in phosphoryl transfer reactions, nicotinamide adenine dinucleotide and its 3'-phosphate (NAD and NADP) in oxidoreductions, acetyl CoA in acyl transfer reactions, 3'-phosphoadenosine-5'-phosphosulphate (PAPS) in the formation of sulphate esters, and sugar nucleotides in glycosyl transfer reactions. A method for regeneration of S-adenosyl methionine in enzymatic methyl transfer reactions has yet to be attained.

In combination with regeneration systems, many cofactor-dependent reactions have been applied on preparative or industrial scales. Fig. 3a illustrates *in situ* sugar nucleotide synthesis in conjunction with enzymatic glycosylation for the assembly of oligosaccharides⁵. Similarly, recycling of the sulphotransferase cofactor PAPS has been used for the conversion of oligosaccharides to their sulphate-containing analogues (Fig. 3b)²². Coupling glycosyl- and sulphotransferase cofactor regeneration cycles may lead to facile preparation of sulphated oligosaccharides or glycopeptides. In this regard, the tyrosine-sulphate-containing glycopeptide from the amino terminus of P-selectin glycoprotein ligand-1 (PSGL-1) represents a potential biologically relevant target structure (Fig. 3c). In addition, one-pot processes using two or three glycosyltransferases coupled with cofactor regeneration have been demonstrated for the

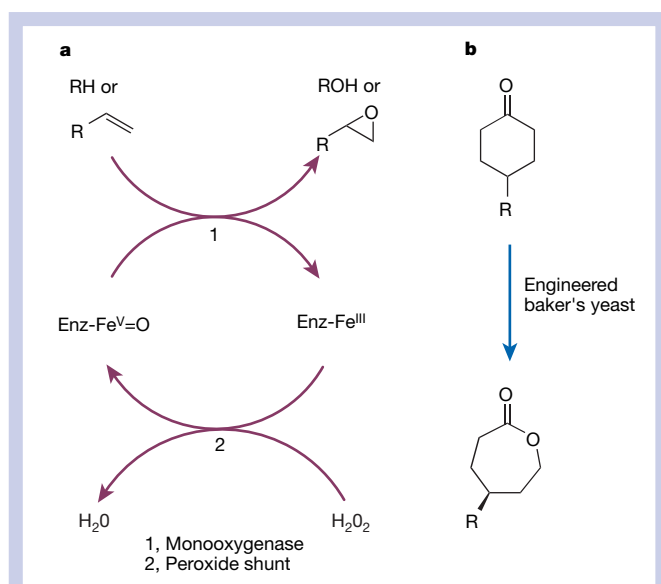


Figure 4 Monooxygenase-catalysed reaction. **a**, Use of the peroxide-shunt pathway in a monooxygenase reaction to avoid regeneration of the cofactor NADPH.

b, Engineered baker's yeast containing a cyclohexanone monooxygenase for asymmetric Baeyer–Villiger oxidation.

synthesis of complex oligosaccharides⁵, including sialyl Lewis x (ref. 5), the sialyl-T antigen²³ and a hyaluronic acid polymer²⁴.

Reduction of the number of enzymes required for cofactor recycling has been accomplished by expressing two enzymatic activities as a single protein, yielding bifunctional fusion enzymes. These bifunctional biocatalysts improve the efficiency of enzymatic reactions, as well as the stability of the enzymes. Examples include the fusion proteins cytidine-5'-monophosphate (CMP)-sialic acid synthetase/ α 2,3-sialyltransferase²⁵ and uridine diphosphate (UDP)-glucose epimerase/ α 1,3-galactosyltransferase²⁶. Alternatively, certain cofactors can be generated and regenerated *in vivo* in cell-based processes. Through metabolic engineering, certain sugar nucleotide-dependent glycosyltransferases (so far, α 1,4-galactosyltransferase and α 2,3-sialyltransferase) have been used in coupled bacterial cell-based synthesis of oligosaccharides²⁷. These systems are notable for producing complex oligosaccharides from simple and inexpensive monosaccharide building blocks. Recent advances in sugar nucleotide-dependent syntheses include the use of polyphosphate kinase for uridine 5'-triphosphate (UTP) regeneration²⁸, the preparation of unnatural oligosaccharides^{29–31} and the synthesis of the *N*-glycan core trisaccharide³². The regeneration principle and strategy should also be applicable to enzymatic syntheses using different sugar nucleotides, such as thymidine diphosphate (TDP) sugars³³. Complex oligosaccharides are difficult to obtain through strictly chemical synthesis, and biocatalytic methods for their construction have aided evaluation of the biological significance of carbohydrates.

Many cofactor-dependent oxygenases, including mono- and dioxxygenases, activate molecular oxygen (O₂), and insert an oxygen atom stereoselectively into unreactive molecules such as alkanes, aromatics and olefins. This catalytic process has great potential in synthetic chemistry, but is difficult to achieve with non-biological methods. Oxygenases use the low oxidation state of metals (such as Fe²⁺ or Cu⁺) for activation of oxygen, and their regeneration often requires NAD(P)H. For synthetic applications, whole cells with regeneration of NAD(P)H *in vivo* have been used instead of free enzymes, largely because free enzymes are inactivated by the reactive radical intermediates generated in the reaction process. The reactive oxo-iron species required in the cytochrome P450 monooxygenase reaction can be obtained from the ferric state by reaction with

hydrogen peroxide rather than from the ferrous state by reaction with NADPH and molecular oxygen (Fig. 4). This 'peroxide shunt' pathway has been applied to the regioselective oxidation of aromatics with new variants of P450 created by directed evolution³⁴. Alternatively, NAD(P)H can be replaced with other reducing agents, or alteration of the catalytic pathway can avoid the use of the expensive cofactor. For example, the flavin-dependent cyclohexanone monooxygenase used in asymmetric Baeyer–Villiger oxidations has been expressed in baker's yeast, and the engineered yeast has been used in the enantioselective synthesis of cyclic lactones as chiral building blocks (Fig. 4)³⁵. Enzymatic oxidation remains an important synthetic process, and understanding the detailed mechanism of oxygenases^{36–39} will perhaps help to solve the problem of enzyme instability in air, and lead ultimately to the design of new oxidative catalysts.

Other important oxidative processes developed recently include the coupling of glycolate oxidase and catalase for the synthesis of glyoxalate from glycolate in free-enzyme and whole-cell systems⁴⁰. The haloperoxidase enzymes are stable and active in the free form, and have been used as catalysts in the stereoselective epoxidation of various olefins^{41,42} and the synthesis of halohydrins¹⁰. Crosslinked crystals⁴³ of the peroxidase seleno-subtilisin have been used for the resolution of racemic hydroperoxides⁴⁴, and other peroxidases have also been used for phenolic coupling⁴⁵. Other oxygenases also have synthetic utility. Peptide amides have been obtained through peptidylglycine amidating monooxygenase catalysis⁴⁶. Dioxxygenases have been used for the oxidation of aromatics⁴⁷, and α -hydroxyacids have been prepared from α -haloacids using L-2-haloacid dehalogenase⁴⁸.

Proteases and glycohydrolases as synthetic catalysts

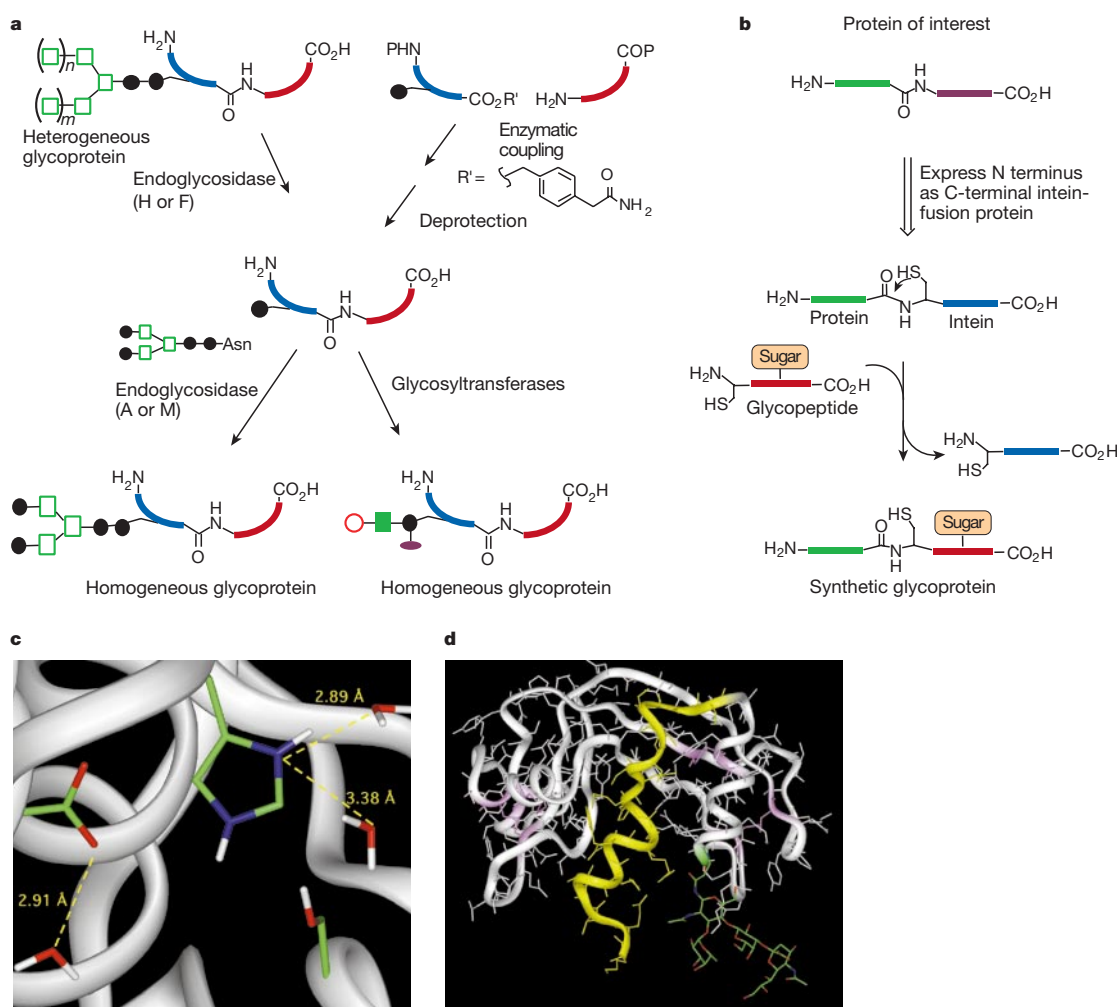
Proteases continue to be important catalysts for peptide synthesis in thermodynamically controlled processes (condensation of the acid and amine groups; reaction 5 in Box 1) or kinetically controlled processes (aminolysis of an ester; reaction 6 in Box 1). Although proteases hydrolyse peptide bonds *in vivo*, these enzymes can be induced to serve as amide- or ester-bond-forming catalysts under specific conditions *in vitro*. Both chemical and genetic methods have been applied to modify the specificity of proteases, and these techniques have led to a mechanistic understanding of various protease-catalysed reactions. In the synthetic direction, the substrate–mimic donor strategy (that is, appending a specific functional group to an acylated donor substrate) allows peptide bonds to be formed between amino acids other than those usually accepted by the protease⁴⁹. The amidase and esterase activities of serine proteases can be modulated by active-site residue modification using chemical methods in conjunction with site-directed mutagenesis⁵⁰. In this scenario, a single amino-acid substitution (Ser to Cys) in the subtilisin active site, previously reported by Kaiser *et al.* (see citation in ref. 51), results in marked changes in enzyme specificity, allowing aminolysis reactions to be carried out effectively^{51,52}.

Addition of water-miscible organic solvent such as dimethylformamide (DMF) to subtilisin BPN' reactions improves the aminolysis

Table 2 Current research activities in enzymatic synthesis

• Development of new enzymatic reactions using designed substrates	• Use of recombinant DNA for large-scale production of enzymes
• Use with organic solvents and co-solvents; one- and two-phase systems	• Mutagenesis and directed evolution for changes in enzymatic properties
• Overcoming substrate and product inhibition	• Exploration of new enzymes and enzymes from new species for synthetic use
• Cofactor regeneration: SAM (to be solved), NAD(P), ATP, CoA, PAPS, sugar nucleotides	• Scale-up in fine chemical and pharmaceutical synthesis
• Modification of enzyme activity and semisynthetic enzymes	• Multienzyme systems for synthesis <i>in vitro</i> and <i>in vivo</i>
• Enzyme immobilization and stabilization	• Antibody catalysis

Figure 5 Enzymatic synthesis of glycoproteins. **a**, N-protected glycopeptide esters can be prepared on solid-phase for segment condensation using engineered serine proteases such as subtilisin, followed by incorporation of additional sugars using glycosyltransferases. Endoglycosidases can also be used to remodel the carbohydrate moiety of heterogeneous glycoproteins through glycosidic cleavage or transglycosylation. **b**, Intein-mediated synthesis of glycoprotein. **c**, The active-site structure of subtilisin BPN' in 50% DMF indicates imidazole flipping and disruption of the H bond between Asp and His. **d**, Model of a sialyl Lewis x-containing ribonuclease A prepared by the method described in **a** (ref. 51).



process and suppresses hydrolysis. X-ray structural investigation of subtilisin BPN' has revealed that the active-site His residue flips in 50% DMF, and the strong hydrogen bond observed between His and Asp in water is disrupted⁵³. This finding was confirmed by nuclear magnetic resonance, and provides a mechanistic rationale for the preferred aminolysis reaction in the DMF–water co-solvent (Fig. 5). Site-directed mutagenesis and directed evolution have also been used for improvement of thermal and/or solvent stability. The development of effective proteases for protein synthesis have produced a subtiligase which served as a peptide bond-forming catalyst in the total synthesis of the intact ribonuclease A protein⁵². Engineered proteases for protein synthesis can provide facile access to proteins that contain unnatural amino acids or functionality, yielding new structures that cannot be obtained through normal biosynthetic pathways.

In addition, engineered subtilisins with improved stability and altered substrate specificity are suitable for glycopeptide coupling, furnishing a route towards glycoprotein synthesis. Glycoproteins produced in nature often present many different glycans⁵¹. This makes analysis of glycan structure, as well as its effect on underlying protein structure and function, nearly impossible to assess. So new methods for the synthesis of glycoproteins with homogeneous glycoforms are needed for the systematic understanding of glycan function to progress, and for the development of glycoprotein pharmaceuticals. To this end, glycoproteins have been assembled through protease-catalysed coupling of glycopeptide ester segments prepared by solid-phase synthesis, followed by enzymatic incorporation of additional sugars using glycosyltransferases (Fig. 5)⁵¹. Other

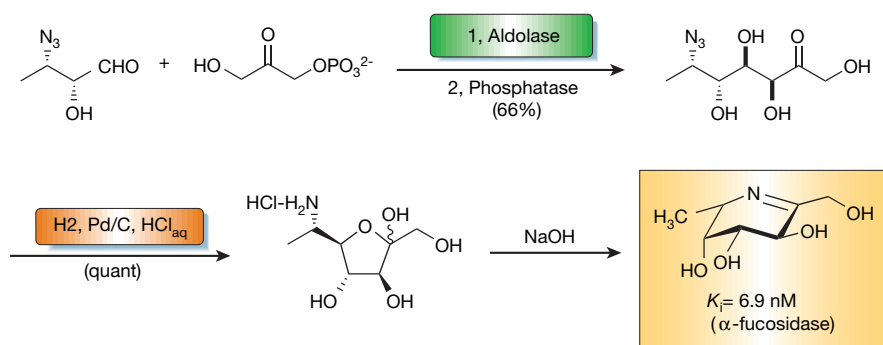
enzymatic methods developed in this regard include the use of inteins for glycoprotein condensation⁵⁴, and the use of endoglycosidases for cleavage or exchange of sugar chains (Fig. 5)⁵⁵. Analogous to proteases, endoglycosidases normally cleave internal glycosidic linkages in an oligosaccharide chain, but can be used as synthetic catalysts under kinetically controlled conditions.

In addition to glycosyltransferases and endoglycosidases, various exoglycosidases (which cleave non-reducing terminal glycosidic linkages *in vivo*) have been applied to the formation of glycosidic bonds. New donor substrates investigated for synthetic transglycosylation reactions include glycosyl fluorides^{56,57}, oxazolines⁵⁸ and 6-oxo-glycosides⁵⁹. Of particular significance is the use of site-directed mutagenesis in the conversion of enzymatic function from exoglycosidase activity to glycosynthase activity^{56,57}. So far, this strategy has been used for the enzymes β -glucosidase and cellulase. In both cases, mutagenesis of the nucleophilic catalytic carboxyl group (Glu) to Ala abolishes hydrolytic activity. The mutant glycosidases are then used with activated glycosyl donors (such as glycosyl fluorides) of the opposite anomeric configuration as the normal substrate for the synthesis of oligosaccharides. Furthermore, the endoglycosidase ceramide glycanase has been used to transfer an oligosaccharyl group from a water-soluble polymer to ceramide. This strategy illustrates an efficient new method for the enzymatic polymer-supported synthesis of glycosphingolipids⁶⁰.

Carbon–carbon bond formation

The construction of C–C bonds with complete stereochemical control is of utmost importance in organic synthesis, and enzyme-catalysed

Figure 6 A representative chemoenzymatic preparation of cyclic imine sugars. The enzyme is fructose-1,6-bisphosphate aldolase.



aldol addition reactions (typically between an aldehyde and a ketone) have made important contributions in this regard¹⁴. In aldolase-catalysed reactions, the enzyme generally controls configuration of newly formed stereogenic centres, although some exceptions in which the substrates affect stereochemical reaction course have been documented. Aldolases are highly specific for the donor substrate (that is, the nucleophilic enolate) but relatively flexible with respect to the acceptor (electrophilic) group. Exploiting this enzymatic characteristic, a judicious choice of acceptor substrate and aldolase has led to the preparation of numerous carbohydrates and mimics thereof¹⁴. These molecules have further served as intermediates in the synthesis of complex bioactive molecules, such as glycosyltransferase and glycosidase inhibitors. The activities of enzymes like fucosidase and fucosyltransferase are implicated in inflammation and in cancer and various other diseases. The inhibition of such enzymes is therefore of potential therapeutic value. A characteristic synthetic example is the use of the dihydroxyacetone phosphate (DHAP)-dependent fructose diphosphate (FDP) aldolase for the construction of cyclic imine sugars as inhibitors of glycosidases and as building blocks for the synthesis of glycosyltransferase inhibitors (Fig. 6)¹⁴. FDP aldolase has also been used for the synthesis of bicyclic sugars and disaccharide mimics⁶¹.

Pyruvate-dependent aldolases from various sources, such as 2-keto-3-deoxy-6-phosphogluconate aldolase (KDPG aldolase), have been used for stereocontrolled carbon–carbon bond formation⁶². Sialic acid aldolase has been used for specific ¹³C labelling of the sialic acid 3-position carbon. After further conversion to [¹³C]-CMP-NeuAc, the labelled sialic acid was transferred to the surface of a glycoprotein for conformational analysis⁶³. The acetaldehyde-dependent aldolase 2-deoxyribose-5-phosphate aldolase (DERA) is the only known aldolase that catalyses condensation between two aldehydes, and has been used in the synthesis of epothilones (Box 1, reaction 7)¹⁴, a new class of anti-cancer agents of

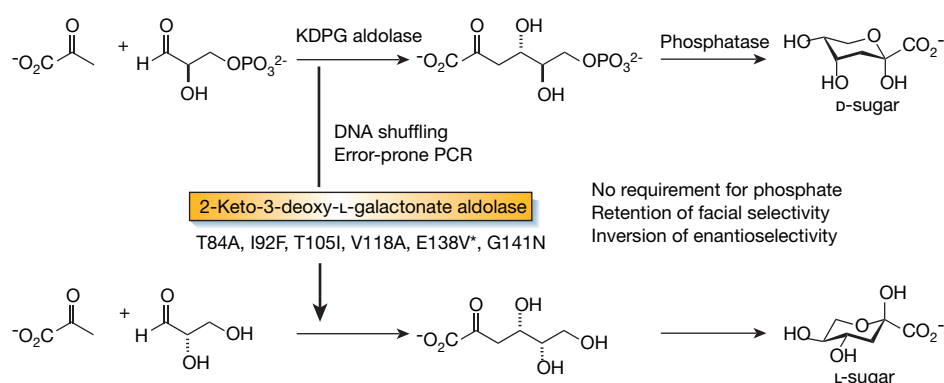
interest in the pharmaceutical industry. The glycine-dependent D- and L-threonine aldolases have provided modified β -hydroxy- α -amino acids that are components of numerous natural products¹⁴. Other synthetically useful enzymes catalysing C–C bond formation include transaldolases, transketolases, cyanohydrin synthetases (also called oxynitrilase) and enzymes for acyloin condensation, acyl-transfer, isoprenoid and steroid assembly, β -replacement of amino acids, and many B₁₂-dependent reactions^{3,4}. For example, DAHP synthetase has been used as a component of metabolically engineered microorganisms for the large-scale production of vanillin from glucose⁶⁴. Furthermore, catalytic antibodies developed in recent years have the ability to match the efficiency of the natural aldolases, while accepting a more diverse range of substrates⁶⁵. A practical synthetic application of the catalytic antibody aldolase Ab 38C2 was illustrated in the synthesis of epothilones (Box 1, reaction 8).

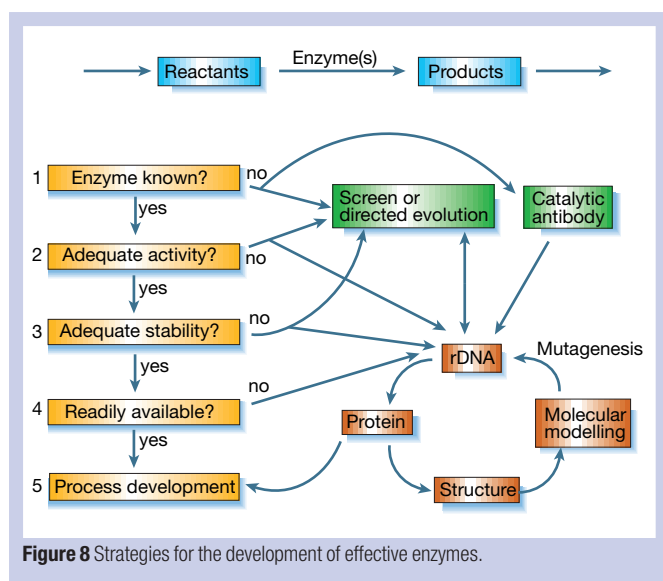
Recent advances and future development

Over the past twenty years, protein engineering based on site-directed mutagenesis has contributed significantly to our understanding of enzyme catalysis, and has led to the development of enzyme variants with modified properties for synthetic transformations. This rational approach has experienced only limited success, however, with regard to creating new enzymes for synthetic applications. One of the main hindrances associated with the rational manipulation of protein primary sequence is the longstanding inability to predict the exact protein structure required for the stereoselective reaction of a given substrate.

Recently, *in vitro*-directed evolution of enzymes, using random genetic mutation and recombination, followed by screening or selection for a desired trait, has been explored as a more generally applicable approach to the modification of enzyme properties (see the review by Arnold in this issue, pages 253–257)^{66,67}. This technique

Figure 7 Directed evolution of 2-keto-3-deoxy-D-phosphogluconate (KDPG) aldolase (which is highly specific for D-glyceraldehyde-3-phosphate) to a new aldolase variant effectively accepting both D- and L-glyceraldehyde to make D- and L-sugars.





has the advantage that it does not require *a priori* knowledge of the relationship between protein structure and function for experimental design. Current progress has shown that directed evolution can yield new enzymes with altered substrate specificity, enantioselectivity, protein topology, thermal stability and tolerance to organic solvents. For example, lipase from the ubiquitous environmental bacterium *Pseudomonas aeruginosa* was evolved to catalyse the hydrolysis of a model ester with >90% enantiomeric excess, compared with 2% enantiomeric excess for the wild-type enzyme⁶⁷. A cytochrome P450 monooxygenase from *P. putida* was evolved to catalyse the hydroxylation of naphthalene using hydrogen peroxide with more than 20-fold higher activity than the native enzyme³⁴. The *Escherichia coli* KDPG aldolase, which is highly dependent on phosphate and D-sugars, was evolved to a new variant capable of accepting both D- and L-substrates (Fig. 7)⁶⁸. This new enzyme, which lacks the phosphate requirement, has been exploited for the synthesis of D- and L-sugars. Notably, the six mutations found in this variant are not in the active site, underscoring the power of directed evolution and the unpredictable factors that influence enzyme specificity.

Phage-display selection methods (the phage-display method originally developed by Smith; see citation in refs 69, 70) also serve to endow proteins with new catalytic activities^{69,70}. Through phage-capture techniques, the selection of new or improved enzymes, as well as catalytic antibodies, is possible. Variations of these methods currently being pursued involve capturing the phage through binding a substrate. This generates either a tagged product or a reactive product that is covalently linked to the phage.

Recent advances in molecular genetics allow the modification of cellular biochemistry to redirect metabolic or biosynthetic pathways, a process called metabolic engineering⁷¹. This new biocatalytic method has been used for the production of many primary and secondary metabolites and their analogues, including aromatics, polyhydroxyalkanoates, antibiotics, polyketides and non-ribosomal peptides (see the reviews in this issue by Khosla and Harbury, pages 247–252, and Walsh, pages 226–231).

The above-mentioned biological tools, together with traditional screening^{72–74} are now available to the chemist to aid in the development of new enzymes for chemical synthesis. Table 2 summarizes current research in enzymatic synthesis, and Fig. 8 provides a strategic approach to the development of enzymatic catalysts for reactions of interest. To select an enzyme for a given reaction, one can start with one enzyme capable of catalysing that specific type of reaction, optimize the reaction conditions, and further improve the catalyst through directed evolution and the protein engineering cycle. In the case where there is no known enzyme for the desired reaction,

non-biological methodology may be the method of choice. Alternatively, approaches based on screening for new enzymes and catalytic antibodies can be pursued if the reaction is sufficiently important. □

- Jones, J. B. Enzymes in organic synthesis. *Tetrahedron* **42**, 3351–3403 (1986).
- Sih, C. J. & Wu, S. H. Resolution of enantiomers via biocatalysis. *Topics Stereochem.* **19**, 63–125 (1989).
- Wong, C.-H. & Whitesides, G. M. *Enzymes in Synthetic Organic Chemistry* (Pergamon, Oxford, 1994).
- Drauz, K. & Waldmann, H. *Enzyme Catalysis in Organic Synthesis* (Wiley, Weinheim, 1995).
- Wong, C.-H., Halcomb, R. L., Ichikawa, Y. & Kajimoto, T. Enzymes in organic synthesis: application to the problems of carbohydrate recognition. Parts 1 and 2. *Angew. Chem. Int. Edn Engl.* **34**, 412–432; 521–546 (1995).
- Schoffers, E., Golebiowski, A. & Johnson, C. R. Enantioselective synthesis through enzymatic asymmetric. *Tetrahedron* **52**, 3769–3826 (1996).
- Faber, K. *Biotransformations in Organic Chemistry* 3rd edn (Springer, Berlin, 1997).
- Klibanov, A. M. Why are enzymes less active in organic solvents than in water? *Trends Biotechnol.* **15**, 97–101 (1997).
- Schmid, R. D. & Verger, R. Lipases: interfacial enzymes with attractive applications. *Angew. Chem. Int. Edn Engl.* **37**, 1608–1633 (1998).
- Zaks, A. & Dodds, D. R. Biotransformations in the discovery and development of pharmaceuticals. *Curr. Opin. Drug. Discov.* **1**, 290–303 (1998).
- Roberts, S. M. *Biocatalysis for Fine Chemicals Synthesis* (Wiley, Weinheim, 1999).
- Sugai, T. Application of enzyme- and microorganism-catalyzed reactions to organic synthesis. *Curr. Org. Chem.* **3**, 373–406 (1999).
- Carrea, G. & Riva, S. Properties and synthetic applications of enzymes in organic solvents. *Angew. Chem. Int. Edn Engl.* **39**, 2226–2254 (2000).
- Machajewski, T. D. & Wong, C.-H. The catalytic asymmetric aldol reaction. *Angew. Chem. Int. Edn Engl.* **39**, 1352–1374 (2000).
- Fang, C.-M. & Wong, C.-H. Enzymes in organic synthesis: alteration of reversible reactions to irreversible processes. *Synlett* **6**, 393–402 (1994).
- Kazlauskas, R. J. Molecular modeling and biocatalysis: explanations, predictions, limitations, and opportunities. *Curr. Opin. Chem. Biol.* **4**, 81–88 (2000).
- Um, P.-J. & Drueckhammer, D. G. Dynamic enzymatic resolution of thioesters. *J. Am. Chem. Soc.* **120**, 5605–5610 (1998).
- Persson, B. A., Larsson, A. L. E., Le Ray, M. & Backvall, J.-E. Ruthenium- and enzyme-catalyzed dynamic kinetic resolution of secondary alcohols. *J. Am. Chem. Soc.* **121**, 1645–1650 (1999).
- Takayama, S., Lee, S. T., Hung, S.-C. & Wong, C.-H. Designing enzymatic resolution of amines. *Chem. Commun.* 127–128 (1999).
- Pathak, T. & Waldmann, H. Enzymes and protecting group chemistry. *Curr. Opin. Chem. Biol.* **2**, 112–120 (1998).
- Bader, B. *et al.* Bioorganic synthesis of lipid-modified proteins for the study of signal transduction. *Nature* **403**, 223–226 (2000).
- Kren, V. & Thiem, J. A multienzyme system for a one-pot synthesis of sialyl T-antigen. *Angew. Chem. Int. Edn Engl.* **34**, 893–895 (1995).
- Burkart, M. D., Izumi, M. & Wong, C.-H. Enzymatic regeneration of 3'-phosphoadenosine-5'-phosphosulfate using aryl sulfotransferase for the preparative enzymatic synthesis of sulfated carbohydrates. *Angew. Chem. Int. Edn Engl.* **38**, 2747–2750 (1999).
- DeLuca, C. *et al.* Enzymatic synthesis of hyaluronic acid with regeneration of sugar nucleotides. *J. Am. Chem. Soc.* **117**, 5869–5870 (1995).
- Gilbert, M. *et al.* The synthesis of sialylated oligosaccharides using a CMP-Neu5Ac synthetase/sialyltransferase fusion. *Nature Biotechnol.* **16**, 769–772 (1998).
- Wang, J.-Q. *et al.* Enhanced inhibition of human anti-gal antibody binding to mammalian cells by synthetic α -gal epitope polymers. *J. Am. Chem. Soc.* **121**, 8174–8181 (1999).
- Endo, T., Koizumi, S., Tabata, K. & Ozaki, A. Large-scale production of CMP-NeuAc and sialylated oligosaccharides through bacterial coupling. *Appl. Microbiol. Biotechnol.* **53**, 257–261 (2000).
- Noguchi, T. & Shiba, T. Use of *Escherichia coli* polyphosphate kinase for oligosaccharide synthesis. *Biosci. Biotechnol. Biochem.* **62**, 1594–1596 (1998).
- Qian, X., Sujino, K., Otter, A., Palcic, M. M. & Hindsgaul, O. Chemoenzymatic synthesis of α -(1,3)-Gal(NAc)-terminating glycosides of complex tertiary sugar alcohols. *J. Am. Chem. Soc.* **121**, 12063–12072 (1999).
- Nishida, Y., Tamakoshi, H., Kitagawa, Y., Kobayashi, K. & Thiem, J. A novel β -1,4-galactosyltransferase reaction to yield β -D-galactopyranosyl-(1-3)-linked disaccharides from L-sugars. *Angew. Chem. Int. Edn Engl.* **39**, 2000–2003 (2000).
- Tsuruta, O., Shinohara, G., Yuasa, H. & Hashimoto, H. UDP-N-acetyl-5-thio-galactosamine is a substrate of lactose synthase. *Bioorg. Med. Chem. Lett.* **7**, 2523–2526 (1997).
- Watt, G. M., Revers, L., Webberly, M. C., Wilson, I. B. H. & Flitsch, S. L. Efficient enzymatic synthesis of the core trisaccharide of N-glycans with a recombinant β -mannosyltransferase. *Angew. Chem. Int. Edn Engl.* **36**, 2354–2356 (1997).
- Stein, A., Kula, M.-R. & Elling, L. Combined preparative enzymatic synthesis of dTDP-6-deoxy-4-keto-D-glucose from dTDP and sucrose. *Glycoconjugate J.* **15**, 139–145 (1998).
- Joo, H., Lin, Z. & Arnold, F. H. Laboratory evolution of peroxide-mediated cytochrome P450 hydroxylation. *Nature* **399**, 670–673 (1999).
- Stewart, J. D. *et al.* Recombinant baker's yeast as a whole-cell catalyst for asymmetric Baeyer-Villiger oxidations. *J. Am. Chem. Soc.* **120**, 3541–3548 (1998).
- Burzlaff, N. I. *et al.* The reaction cycle of isopenicillin N-synthase observed by X-ray diffraction. *Nature* **401**, 721–724 (1999).
- Schlichting, I. *et al.* The catalytic pathway of cytochrome P450cam at atomic resolution. *Science* **287**, 1615–1622 (2000).
- Elliott, S. J. *et al.* Regio- and stereoselectivity of particulate methane monooxygenase from *Methylococcus capsulatus*. *J. Am. Chem. Soc.* **119**, 9949–9955 (1997).
- Rosenzweig, A. C., Nordlund, P., Takahara, P. M., Frederick, C. A. & Lippard, S. J. Geometry of the soluble methane monooxygenase catalytic diiron center in two oxidation states. *Chem. Biol.* **2**, 409–418 (1995).

40. Gavagan, J. E. *et al.* Glyoxylic acid production using microbial transformant catalysts. *J. Org. Chem.* **60**, 3957–3963 (1995).
41. Lakner, F. J., Cain, K. P. & Hager, L. P. Enantioselective epoxidation of ω -bromo-2-methyl-1-alkenes catalyzed by chloroperoxidase. Effect of chain length on selectivity and efficiency. *J. Am. Chem. Soc.* **119**, 443–444 (1999).
42. van Deurzen, M. P., van Rantwijk, F. & Sheldon R. A. Selective oxidation catalyzed by peroxidases. *Tetrahedron* **53**, 13183–13220 (1997).
43. Margolia, A. L. Cross-linked enzyme crystals as novel materials for catalysis and chromatography. *Chimia* **50**, 670–673 (1999).
44. Haring, D. & Schreier, P. Novel biocatalysts by chemical modification of known enzymes: cross-linked microcrystals of the semisynthetic peroxidase seleno-subtilisin. *Angew. Chem. Int. Edn Engl.* **37**, 2471–2473 (1998).
45. Guo, Z.-w., Salamonczyk, G. M., Han, K., Machiya, K. & Sih, C. J. Enzymatic oxidative phenolic coupling. *J. Org. Chem.* **62**, 6700–6701 (1997).
46. Prigge, S. T., Kolhekar, A. S., Eipper, B. A., Mains, R. E. & Amzel, L. M. Amidation of bioactive peptides: the structure of peptidylglycine α -hydroxylating monooxygenase. *Science* **278**, 1300–1305 (1997).
47. Hudlicky, T. *et al.* Toluene dioxygenase-mediated *cis*-hydroxylation of aromatics in enantioselective synthesis. Asymmetric total syntheses of pancratistatin and 7-deoxypancratistatin, promising antitumor agents. *J. Am. Chem. Soc.* **118**, 10752–10765 (1996).
48. Liu, J.-Q., Kurihara, T., Miyagi, M., Esaki, N. & Soda, K. Reaction mechanism of L-2-haloacid dehalogenase of *Pseudomonas* sp. YL. *J. Biol. Chem.* **270**, 18309–18312 (1995).
49. Bordusa, F., Ullman, D., Elsner, C. & Jakubke, H.-D. Substrate mimetic mediated peptide synthesis: an irreversible ligation strategy that is independent of substrate specificity. *Angew. Chem. Int. Edn Engl.* **36**, 2473–2475 (1997).
50. Plettner, E., DeSantis, G., Stabile, M. & Jones, J. B. Modulation of esterase and amidase activity of subtilisin *Bacillus lentus* by chemical modification of cysteine mutants. *J. Am. Chem. Soc.* **121**, 4977–4981 (1999).
51. Witte, K., Seitz, O. & Wong, C.-H. Solution- and solid-phase synthesis of N-protected peptide esters of the benzyl type as substrates for subtilisin-catalyzed glycopeptide coupling. *J. Am. Chem. Soc.* **120**, 1979–1989 (1998).
52. Jackson, D. Y. *et al.* A designed peptide ligase for total synthesis of ribonuclease A with unnatural catalytic residues. *Science* **266**, 243–247 (1994).
53. Kidd, R. D. *et al.* Breaking the low barrier hydrogen bond in a serine protease. *Protein Sci.* **8**, 410–417 (1999).
54. Tolbert, T. J. & Wong, C.-H. Intein-mediated synthesis of proteins containing carbohydrates and other molecular probes. *J. Am. Chem. Soc.* **122**, 5421–5428 (2000).
55. Wang, L.-X. *et al.* Combined chemical and enzymatic synthesis of a C-glycopeptide and its inhibitory activity toward glycoamidases. *J. Am. Chem. Soc.* **119**, 11137–11146 (1997).
56. Mackenzie, L. F., Wang, Q., Warren, R. A. J. & Withers, S. G. Glycosynthases: mutant glycosidases for oligosaccharide synthesis. *J. Am. Chem. Soc.* **120**, 5583–5584 (1998).
57. Fort, S. *et al.* Highly efficient synthesis of β (1,4)-oligo- and polysaccharides using a mutant cellulase. *J. Am. Chem. Soc.* **122**, 5429–5437 (2000).
58. Kobayashi, S., Kiyosada, T. & Shoda, S.-i. Synthesis of artificial chitin: irreversible catalytic behavior of a glycosyl hydrolase through a transition state analogue substrate. *J. Am. Chem. Soc.* **118**, 13113–13114 (1996).
59. Kimura, T., Takayama, S., Huang, H. & Wong, C.-H. A practical method for the synthesis of N-acetyl-D-lactosamine derivatives by the tandem use of galactose oxidase and β -galactosidase. *Angew. Chem. Int. Edn Engl.* **35**, 2348–2350 (1996).
60. Nishimura, S.-I. & Yamada, K. Transfer of ganglioside GM3 oligosaccharide from a water soluble polymer to ceramide by ceramide glycanase. A novel approach for the chemical-enzymatic synthesis of glycosphingolipids. *J. Am. Chem. Soc.* **119**, 10555–10556 (1997).
61. Zannetti, M. T., Walter, C., Knorst, M. & Fessner, W.-D. Fructose-1,6-bisphosphate aldolase from *Staphylococcus carnosus*: overexpression, structure prediction, stereoselectivity, and application in the synthesis of bicyclic sugars. *Chem. Eur. J.* **5**, 1882–1890 (1999).
62. Shelton, M. C. *et al.* 2-Keto-3-deoxy-6-phosphogluconate aldolases as catalysts for stereocontrolled carbon-carbon bond formation. *J. Am. Chem. Soc.* **118**, 2118–2125 (1996).
63. Miyazaki, T., Sato, H., Sakakibara, T. & Kajihara, Y. An approach to the precise chemoenzymatic synthesis of ^{13}C -labeled sialyloligosaccharide on an intact glycoprotein: a novel one-pot $[3-^{13}\text{C}]$ -labeling method for sialic acid analogues by control of the reversible aldolase reaction, enzymatic synthesis of $[3-^{13}\text{C}]$ -NeuAc- α -(2,3)-[U- $^{13}\text{C}]$ -Gal- β -(1,4)-GlcNAc- β - sequence onto glycoprotein, and its conformational analysis by developed NMR techniques. *J. Am. Chem. Soc.* **122**, 5678–5694 (2000).
64. Li, K. & Frost, J. W. Synthesis of vanillin from glucose. *J. Am. Chem. Soc.* **120**, 10545–10546 (1998).
65. Hoffmann, T. *et al.* Aldolase antibodies of remarkable scope. *J. Am. Chem. Soc.* **120**, 2768–2779 (1998).
66. Arnold, F. H. Design by directed evolution. *Acc. Chem. Res.* **31**, 125–131 (1998).
67. Reetz, M. T. & Jaeger, K.-E. Enantioselective enzymes for organic synthesis created by directed evolution. *Chem. Eur. J.* **6**, 407–412 (2000).
68. Fong, S., Machajewski, T. D., Mak, C. C. & Wong, C.-H. Directed evolution of D-2-keto-3-deoxy-6-phosphogluconate aldolase to new variants for the efficient synthesis of D- and L-sugars. *Chem. Biol.* **7**, 873–883 (2000).
69. Janda, K. D. *et al.* Chemical selection for catalysis in combinatorial antibody libraries. *Science* **275**, 945–948 (1997).
70. Jestin, J.-L., Kristensen, P. & Winter, G. A method for the selection of catalytic activity using phage display and proximity coupling. *Angew. Chem. Int. Edn Engl.* **38**, 1124–1127 (1999).
71. Cane, D. E., Walsh, C. T. & Khosla, C. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science* **282**, 63–68 (1998).
72. Shimizu, S., Ogawa, J., Kataoka, M. & Kobayashi, M. Screening of novel microbial enzymes for the production of biologically and chemically useful compounds. *Adv. Biochem. Eng. Biotechnol.* **58**, 45–88 (1997).
73. Schoemaker, H. E. *et al.* Application of enzymes in industrial organic synthesis. *Chimia* **51**, 308–310 (1997).
74. Patel, R. N. *Stereoselective Biocatalysis* (Dekker, New York, 2000).

Acknowledgements

We thank P. Sears for assistance with figure preparation, and other co-workers whose names are listed in the references. We also extend apologies to those whose biocatalysis research was not cited owing to space constraints. This research was supported by the NIH and NSF.

Improving enzymes by using them in organic solvents

Alexander M. Klibanov

Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA (e-mail klibanov@mit.edu)

The technological utility of enzymes can be enhanced greatly by using them in organic solvents rather than their natural aqueous reaction media. Studies over the past 15 years have revealed not only that this change in solvent is feasible, but also that in such seemingly hostile environments enzymes can catalyse reactions impossible in water, become more stable, and exhibit new behaviour such as 'molecular memory'. Of particular importance has been the discovery that enzymatic selectivity, including substrate, stereo-, regio- and chemoselectivity, can be markedly affected, and sometimes even inverted, by the solvent. Enzyme-catalysed reactions in organic solvents, and even in supercritical fluids and the gas phase, have found numerous potential applications, some of which are already commercialized.

The tremendous potential of enzymes as practical catalysts is well recognized^{1,2}. In particular, they are being increasingly exploited for asymmetric synthetic transformations³, fuelled by the growing demand for enantiopure pharmaceuticals⁴. But as long as the use of enzymes is restricted to their natural, aqueous reaction media, the scope of industrial bioconversions, especially for the production of speciality chemicals and polymers, is necessarily limited by a variety of considerations. Most such compounds are insoluble in water, and water frequently gives rise to unwanted side reactions and degrades common organic reagents. The thermodynamic equilibria of many processes are unfavourable in water, and product recovery is sometimes difficult from this medium.

In principle, most of these problems might be overcome by switching from water to organic solvents as the reaction media. At first sight, this substitution would seem impossible in the light of the conventional idea that enzymes (and other proteins) are denatured (lose their native structure and thus catalytic activity) in organic solvents⁵. This notion, however, comes from examining enzymes in aqueous–organic mixtures⁶, not in neat (pure) organic solvents. Although it is tempting to assume that if enzymes denature in the former medium, they will certainly suffer the same fate in the latter, this assumption has now been shown to be wrong⁷. The reason for this counter-intuitive behaviour is that in the absence of water, which acts as a molecular lubricant^{8,9}, enzymes are very rigid. Consequently, although in aqueous–organic mixtures protein molecules have both a proclivity to denature and sufficient conformational flexibility to do so, in dry solvents their drive to unfold is greater still but the pliability necessary to proceed is lacking⁷. As a result, various crystalline enzymes essentially retain their native structures even in anhydrous organic solvents^{10–14}.

It is, then, perhaps not so startling that studies over the past 15 years have established firmly that many enzymes can work in organic solvents containing little or no water^{15,16}. Surprises have, and probably will continue to, come from discovering new, unique and useful properties that enzymes exhibit in such seemingly unlikely and hostile media. Here I review some of these properties.

Enzymatic activity in organic solvents

The absence of water is in itself often immediately conducive

to new enzymatic reactions. For instance, in water numerous lipases, esterases and proteases catalyse the hydrolysis of esters to the corresponding acids and alcohols. In anhydrous solvents, this process obviously cannot occur. However, adding alternative nucleophiles, such as alcohols, amines and thiols, leads to transesterification, aminolysis and thioesterification, respectively — reactions suppressed in aqueous solution¹⁷. Moreover, the synthesis of esters from their constituent acids and alcohols (the reverse of hydrolysis) becomes thermodynamically favourable¹⁷.

In general, the catalytic activity displayed by enzymes in neat organic solvents is far lower than in water¹⁸. But there may be nothing inevitable about this decline, and both its underlying causes¹⁹ and effective remedies are emerging¹⁸. Hydrophobic solvents are usually superior to hydrophilic ones as enzymatic reaction media because the latter have a greater tendency to strip tightly bound water (which is essential for catalytic activity) from the enzyme molecules²⁰. Also, because proteins are insoluble in almost all organic solvents, enzyme powder suspensions in them should be stirred or shaken vigorously to eliminate mass-transfer barriers for substrates. The powders are usually lyophilized or freeze-dried enzymes. Lyophilization is a gentle dehydration process whereby aqueous solutions are frozen and then placed in vacuum. The water (ice) vaporizes without melting, and the non-volatile components are left behind in a solid, typically undamaged state.

One of the most influential parameters affecting enzymatic activity in aqueous solution is pH. But it has no meaning in organic solvents. Instead, it has been found that enzymes in such media have a 'pH memory': their catalytic activity reflects the pH of the last aqueous solution to which they were exposed^{17,20}. This phenomenon is due to the fact that protein ionogenic groups retain their last ionization state on both dehydration and subsequent placement in organic solvents. Consequently, the enzymatic activity in such media can be much enhanced, sometimes hundreds of times, if enzymes are lyophilized from aqueous solutions of the pH optimal for catalysis^{17,20,21}. Alternatively, the ionization status in organic solvents can be optimized, and hence the enzymatic activity maximized, by adding appropriate buffer pairs of acids and their conjugated bases^{21,22}.

A priori, one might worry that when an enzyme is exposed to an organic solvent its denaturation should ensue. But this does not happen to either crystalline^{10–14} or

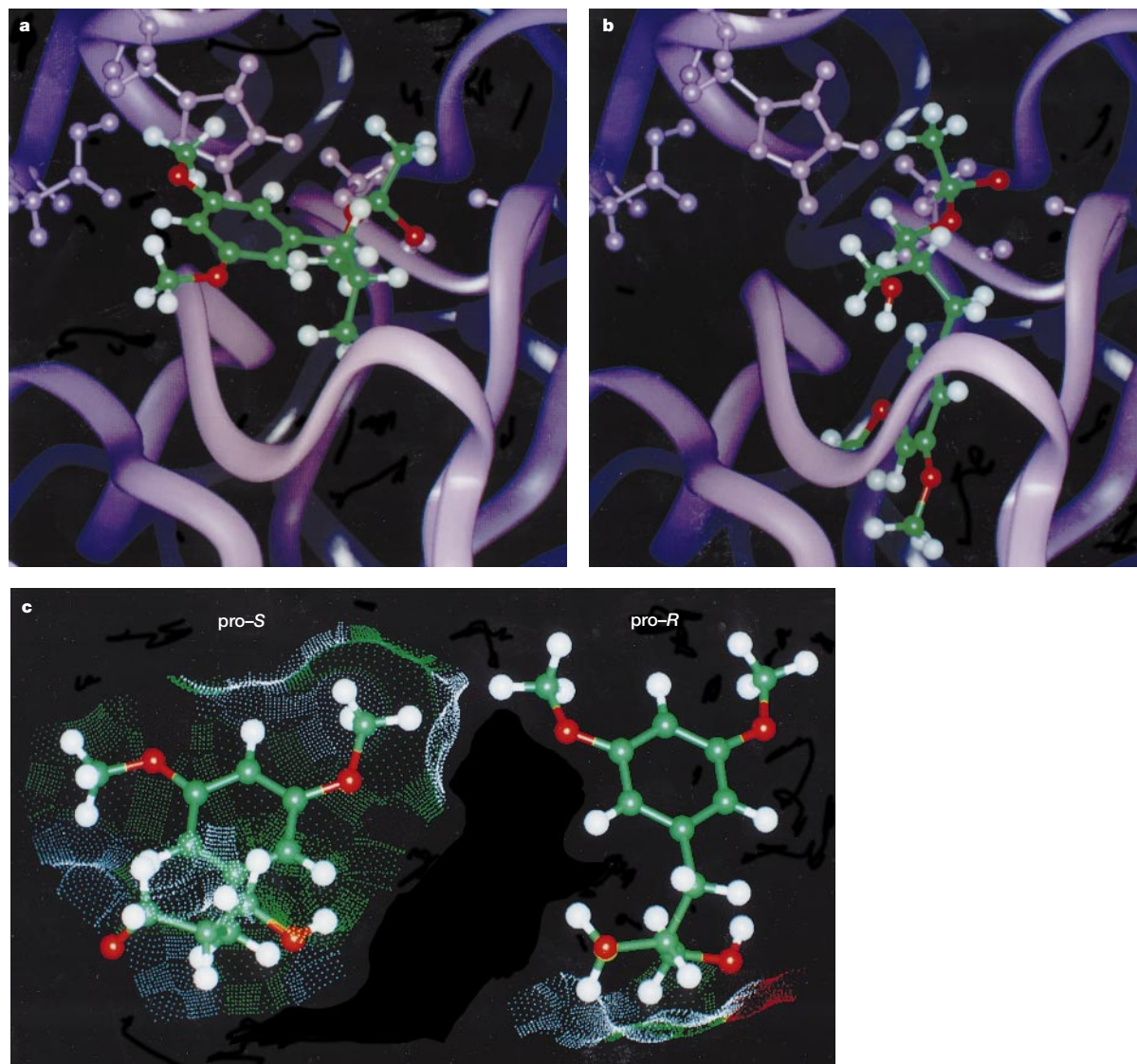


Figure 1 Structural modelling of enzymatic transition states. **a**, **b**, Molecular models and **c**, solvent-accessible surface areas of the α -chymotrypsin-bound pro-*S* (**a**) and pro-*R* (**b**) transition states of 2-(3,5-dimethoxybenzyl)-1,3-propanediol in its enzymatic acetylation⁵². The main chain of chymotrypsin in the active-site region is depicted as a ribbon diagram (in **a** and **b**); the substrate is represented by a ball-and-stick model (green, red and white balls correspond to carbon, oxygen and hydrogen atoms, respectively). In **c**, the dots demarcate the solvent-accessible surfaces of the substrate. Molecular modelling and dynamics simulations were performed using Biosym's Insight II and Discover programs; for details, see ref. 52. Note that unlike in the pro-*S* state, the pro-*R* transition state adopts a conformation in which the dimethoxyphenyl group of the substrate is buried in the binding pocket of the enzyme.

lyophilized²³ enzymes. In both instances the anhydrous environment presumably locks the enzyme molecule kinetically in its prior conformation. The lyophilization step can, however, itself cause significant denaturation^{23,24}. In other words, ironically it is not contact with an organic solvent but the prior dehydration that changes the enzyme structure and results in diminished enzymatic activity in organic solvents. This detrimental effect can be greatly minimized or even prevented by lyophilizing enzymes in the presence of structure-preserving lyoprotectants, such as sugars and poly(ethylene glycol)^{25,26}, certain inorganic salts²⁷, substrate-resembling ligands^{25,26,28} and crown ethers²⁹. Another possibility is to form organic-soluble complexes of enzymes with lipids^{30,31}, in which they apparently remain in native, enzymatically active conformations. These approaches have resulted in the activation of lyophilized enzymes in organic solvents by up to four orders of magnitude¹⁸. Alternatively, crystalline enzymes, which are far more resistant to dehydration-caused denaturation, can be used^{19,21,32}.

Yet another important reason for diminished enzymatic activities in organic solvents stems from reduced structural flexibility. In aqueous environments, enzymes possess the conformational mobility necessary for optimal catalysis^{8,9}. In contrast, organic solvents lack water's ability to engage in multiple hydrogen bonds³³, and also have lower dielectric constants, leading to stronger intra-protein electrostatic interactions. Consequently, enzyme molecules are much more rigid^{34,35}. Addition of small quantities of water to enzyme suspensions in anhydrous solvents^{36,37} or raising the thermodynamic activity of water by other means³⁸ can increase the enzymatic prowess in such systems by several orders of magnitude. To a certain extent, this activating effect of water can be mimicked by other solvents capable of forming multiple hydrogen bonds, such as glycerol and ethylene glycol³⁶. The same result of 'loosening up' and consequently activating enzymes in anhydrous solvents has been achieved by adding denaturing co-solvents in quantities insufficient to cause full denaturation³⁹.

A few other, less influential factors can also contribute to the observed lower enzymatic activity in organic solvents compared to water^{18,19}. Again, their elucidation often suggests straightforward cures. Applying such strategies systematically can result in markedly more efficient enzymes in organic solvents, with activities sometimes comparable to those in water^{31,36}.

Stability of enzymes in organic solvents

Two types of enzyme instability, such as thermal, should be distinguished. The first is time-dependent, gradual, irreversible loss of enzymatic activity on exposure to high temperatures. The second is heat-induced, cooperative unfolding (usually almost instantaneous and reversible) of enzyme molecules. Water is a pivotal participant in each case, by promoting both the conformational mobility of protein molecules^{8,9} and such major deleterious reactions as deamidation of Asn/Gln residues and hydrolysis of peptide bonds⁴⁰. Hence one would expect that enzymes should be more thermostable in organic solvents than in water.

Indeed, a number of cases of improved stability of enzymes in non-aqueous media against both types of thermal inactivation have been documented. For example, porcine pancreatic lipase⁴¹, ribonuclease⁴² and α -chymotrypsin²⁰ at 100 °C have half-lives of several hours in anhydrous solvents whereas in water they deactivate within seconds at that temperature. Significantly, the enzyme half-life in an organic solvent drops precipitously when the water content is raised^{41–43}.

Likewise, the thermal unfolding (melting) temperature of bovine pancreatic ribonuclease suspended in the anhydrous alkane nonane is 124 °C, whereas that in water is only 61 °C (ref. 42). Understandably, the resistance to thermal unfolding decreases as the water content of the enzyme powder increases. Another mechanistic insight can be gained from the finding⁴² that the thermostabilities of both types of ribonuclease in nonane are the same as for the enzyme powder (of a given water content) simply exposed to air or argon. Therefore, the hydrophobic solvent is essentially inert towards, and has no appreciable interactions with, the enzyme.

These and other similar data indicate that enzymes are predictably extremely thermostable in anhydrous organic solvents owing to their conformational rigidity in the dehydrated state and the absence of the prevalent covalent reactions responsible for irreversible thermal inactivation of enzymes in aqueous solution. It is worth noting that, as in the case of catalytic activity, enzyme stability is enhanced in neat solvents; in aqueous–organic mixtures it is actually much diminished compared with pure water⁶. This is again because in such a mixture enzymes are both driven to denature by the organic component and enabled to succumb to this denaturation by the aqueous one.

In addition, when placed in organic solvents, enzymes become far more stable against another common cause of inactivation in water: proteolysis²⁰. This is because both the enzymes and the offending proteases (excreted by contaminating microorganisms) are insoluble in such media and thus cannot interact.

Solvent dependence of enzyme specificity

Exquisite selectivity is the hallmark of enzymatic catalysis⁴⁴. It is considered to be an intrinsic property of a given enzyme — to alter the selectivity, the enzyme molecule has to be changed, for example by means of site-directed mutagenesis. This notion is correct so long as the enzyme acts in water — that is, when the reaction medium is essentially fixed — but it no longer holds true if enzymatic processes are carried out in organic solvents. Indeed, there are several documented cases in which various types of enzyme selectivity have been changed profoundly on switching from one solvent to another^{45,46}, including substrate, enantiomeric, prochiral, regio- and chemoselectivities. I shall discuss some representative examples of this phenomenon, and possible mechanistic rationales.

Substrate selectivity is manifested in the ability of an enzyme to discriminate between two distinct, albeit structurally similar,

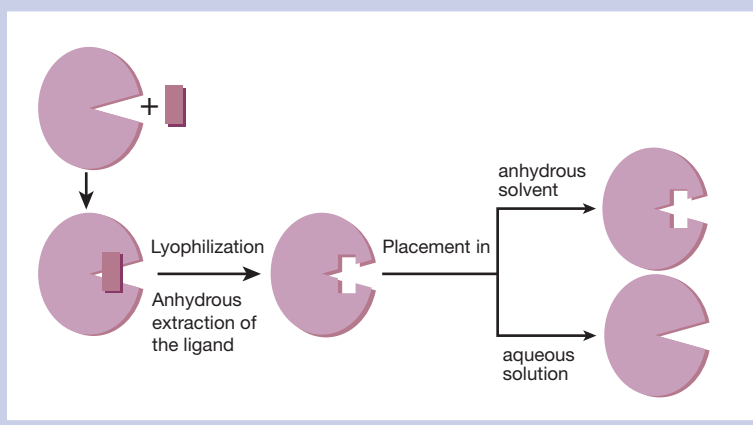
substrates. This is often based on differences in their hydrophobicities. For instance, the main driving force of the enzyme–substrate binding for many proteases, such as α -chymotrypsin and subtilisin, is hydrophobic interactions between the side chain of the amino-acid substrate and the active site of the enzyme⁴⁴. Consequently, a hydrophobic substrate is more reactive than a hydrophilic counterpart simply because this driving force is greater. This situation should change markedly, however, when an organic solvent (in which by definition there are no hydrophobic interactions) is used instead of water. Indeed, we found⁴⁷ that while in water the hydrophobic substrate *N*-acetyl-L-phenylalanine ethyl ester (*N*-Ac-L-Phe-OEt) is some 50,000-fold more reactive than the hydrophilic *N*-acetyl-L-serine ethyl ester (*N*-Ac-L-Ser-OEt) towards α -chymotrypsin, in octane the phenylalanine substrate is about three times less reactive than the serine one. Moreover, whereas subtilisin's reactivity towards *N*-Ac-L-Phe-OEt in dichloromethane is eight times higher than towards *N*-Ac-L-Ser-OEt, exactly the opposite is the case in another organic solvent, *t*-butylamine⁴⁸. Similar pronounced solvent dependences of substrate selectivity have been observed for these two enzymes with other substrates⁴⁹ and with an unrelated enzyme, horseradish peroxidase⁵⁰.

From the synthetic viewpoint, the most valuable type of enzymatic selectivity is stereoselectivity, particularly enantiomeric and prochiral^{1–3}. Unfortunately, enzymes are frequently insufficiently stereoselective in non-natural, practically important transformations, necessitating laborious and time-consuming screening¹. Therefore, the discovery^{45,46} that enzymatic enantio- and prochiral selectivities can be greatly influenced, and sometimes even reversed, by the solvent holds much promise as an alternative to enzyme screening. For instance, the enantioselectivity of α -chymotrypsin in the transesterification of the medicinally important compound methyl 3-hydroxy-2-phenylpropionate with propanol has been found to span a 20-fold range on changing from one organic solvent to another; in fact, whereas the enzyme strongly prefers the *S*-enantiomer of the substrate in some solvents, the *R*-antipode is more reactive in others⁵¹. Likewise, the dominant product of the chymotrypsin-catalysed acetylation of the prochiral substrate 2-(3,5-dimethoxybenzyl)-1,3-propanediol in di-isopropyl ether or cyclohexane is the *S*-monoester, whereas in acetonitrile or methyl acetate the *R*-enantiomer is formed preferentially⁵². Not only do these striking and surprising results seem to be somewhat general^{45,46,53–55}, but at least in some instances they can be explained rationally and almost quantitatively. For example, the solvent-induced variation (and inversion) of chymotrypsin's prochiral selectivity has been rationalized by accounting for the energetics of the substrate desolvation in the enzyme-bound pro-*S* and pro-*R* transition states. Nearly the whole substrate molecule is desolvated in the pro-*R* transition state (Fig. 1), but owing to the steric constraints imposed by the enzyme active site it largely remains solvated in the pro-*S* case. This leads to an entirely different solvent dependence of these two stereochemical reaction pathways and, hence, of prochiral selectivity⁵².

Two other types of enzyme selectivity — regio- and chemoselectivity — are also controlled by the solvent. The former refers to the preference of an enzyme for one out of several identical functional groups in the substrate molecule. It has been shown that such a preference of *Pseudomonas cepacia* lipase for one of the two differently positioned ester groups in an aromatic molecule⁵⁶ or hydroxyl groups in a sugar⁵⁷ is strongly affected (and can be even reversed) by the solvent. Chemoselectivity refers to an extent to which an enzyme favours one of several distinct functional groups in the substrate molecule. For many lipases and proteases, the degree of preference for a hydroxyl group relative to an amino group in a given substrate as an acylation site has been found to depend strongly on the solvent^{58,59}.

Although the profound effect of solvent on various kinds of enzyme selectivity (under otherwise the same conditions) is now firmly proven and ripe for preparative exploitation, it is important to recognize that the underlying mechanisms are only beginning to be elucidated.

Figure 2 Schematic representation of the ligand-induced imprinting of the enzyme active site. The enzyme molecule is depicted as a shaded oval with an angular cleft representing the active site; the ligand molecule is shown as a black rectangle. On binding of the ligand (for example, a substrate analogue) to the enzyme active site in water (denoted by the first arrow), a conformational change occurs, forming an imprint. This altered conformation of the active site remains after lyophilization, followed by extraction of the ligand with a suitable anhydrous solvent (second arrow). In fact, the ligand-induced imprint (memory) persists (top right) even after the enzyme is suspended in an anhydrous solvent, owing to the enzyme's structural rigidity in such media. In contrast, on dissolution in water, where protein molecules are flexible, the imprint disappears (bottom right).



Other features of enzymatic catalysis in organic solvents

One of the most intriguing properties of enzymes in organic solvents is the 'molecular memory' effect⁶⁰ that stems from their high conformational rigidity in anhydrous environments. As a result of this, whereas in water the behaviour of an enzyme is invariant with respect to how its solution has been prepared, enzyme properties in organic media become dependent on history. For example⁶¹, lyophilized α -chymotrypsin first dissolved in water and then diluted 100-fold with *t*-amyl alcohol has a specific activity almost an order of magnitude greater than that of the same lyophilized enzyme directly suspended in that solvent containing the same 1% of water. As extra water is added to this suspension, presumably erasing the memory by making the enzyme pliable, the difference in the enzymatic activities drops.

Furthermore, when subtilisin is lyophilized from aqueous solution containing various competitive inhibitors (followed by their removal by anhydrous extraction), not only is it up to 100 times more active in anhydrous solvents than the enzyme lyophilized in the absence of ligands (owing to the lyoprotection effect), but it also has distinct substrate specificity and stability^{28,62}. This ligand-induced enzyme memory disappears when the enzyme is redissolved in water²⁸. In addition, α -chymotrypsin's enantioselectivity, and the substrate selectivities of lipases, in a given organic solvent are affected markedly by the addition of a ligand to an aqueous solution of the enzyme during dehydration^{63,64}. These findings can be readily explained by assuming that the ligands cause conformational changes in enzyme active sites and that even after the ligand removal such 'imprints' are retained by the enzyme in anhydrous media because of its rigidity in the absence of water (Fig. 2). Since the structures of the ligand-imprinted enzymes are distinct from those of the non-imprinted predecessors, so are the catalytic properties⁶⁰.

Besides molecular imprinting, in some cases conducting enzymatic transformations in organic solvents is beneficial because it offers specific advantages over water. For instance, mandelonitrile lyase catalyses the enantioselective addition of hydrogen cyanide to various aldehydes to give optically active *R*-cyanohydrins⁶⁵. But in aqueous solution, appreciable non-enzymatic addition of hydrogen cyanide occurs, leading to a racemic cyanohydrin, thus compromising the optical yields of the overall process. This spontaneous reaction is suppressed in such organic solvents as ethyl acetate and di-isopropyl ether. Consequently, when they are used as reaction media for biocatalytic cyanohydrin formation, not only is the solubility of the aldehyde substrates greatly increased (leading to higher productivities), but also, owing to the absence of the non-stereoselective chemical reaction, the enantiomeric purity of the product is markedly enhanced⁶⁵.

Another instructive example of the benefits of switching to nonaqueous reaction media also deals with improving enzymatic enantioselectivity, but by entirely different means. It involves a common situation in which the less-reactive substrate enantiomer

experiences greater steric hindrances in the enzyme-bound transition state than the more reactive one. Temporarily enlarging the substrate, for example by forming a salt with a bulky counter-ion, should then exacerbate these hindrances disproportionately for the less-reactive enantiomer and thereby increase the enzymatic enantioselectivity. This strategy would not be viable in water, where salts dissociate into the constituent ions; but such dissociation should not take place in organic solvents. This rationale was verified recently⁶⁶. The enantioselectivity value (*E*)⁶⁷ for the *P. cepacia* lipase-catalysed propanolysis of phenylalanine methyl ester (Phe-OMe) in acetonitrile is merely 5.7, but it jumps to 38 when Phe-OMe's salt with trimethoxycinnamic acid is used as a substrate instead under otherwise identical conditions. Even more strikingly, whereas the enzyme is essentially non-stereoselective in the hydrolysis of 2-benzylsuccinic acid 1-monomethyl ester in *t*-amyl alcohol (*E* = 1.5), it strongly favours the *S*-enantiomer (*E* = 8.1) when a salt with 4-(4-chlorobenzoyl)pyridine is used as a substrate under otherwise equal reaction conditions. Note that the products can be readily recovered from the salts by dissociating them with an acid or a base, and that, as predicted, this enantioselectivity enhancement strategy works only in organic solvents.

Practical applications of enzymes in organic solvents

The use of organic solvents as reaction media can thus greatly expand the repertoire of enzyme-catalysed transformations. Consequently, a number of potential applications of enzymes that are either impossible or marginal in water become quite feasible and commercially attractive in other solvents. Below I consider a few typical and particularly instructive examples involving asymmetric conversions, production of polymers, and analyses.

One of the principal methods for the preparation of optically active acids and alcohols, which are among the most versatile and useful reagents for organic synthesis, has been to esterify a racemic acid or alcohol with an achiral moiety; this is then followed by an asymmetric hydrolysis catalysed by a lipase, esterase or protease (ref. 68, and see reviews in this issue by Walsh, pages 226–231, and Koeller and Wong, pages 232–240). The advent of nonaqueous enzymology¹⁶ allows one to use the same enzymes in direct asymmetric (trans)esterifications, thereby skipping a step (the non-enzymatic esterification required for the subsequent enzymatic hydrolysis) in the resolution process. Following its introduction⁶⁹, this alternative strategy has been explored successfully in hundreds of studies⁷⁰. For instance, enantiopure 2-chloro- and 2-bromo-propionic acids, used as intermediates for the synthesis of phenoxypropionic herbicides and of some pharmaceuticals, have been obtained from yeast lipase-catalysed enantioselective butanolysis in anhydrous solvents. Not only is this process, scaled up by Chemie Linz AG of Austria to a multikilogram level, thermodynamically impractical in water, but water also hinders the resolution by promoting racemization⁶⁹. Schering-Plough, meanwhile, makes hundred-kilogram quantities

of an azole antifungal agent, currently in phase III clinical trials, in a synthetic scheme where a pivotal stereoselective step is the acetylation of a symmetrical diol catalysed by a yeast lipase in acetonitrile³.

Direct asymmetric acylation is even more appealing for preparing chiral amines, as there are far fewer amidase enzymes available compared with lipases and esterases. The last two enzyme types are incapable of hydrolysing amides, but they are able to use amines as nucleophiles, and hence form amide bonds, in organic solvents¹⁷. Such an enzymatic resolution of racemic amines has been validated (and its critical dependence on the solvent established)⁷¹ and scaled up to a kilogram level⁷². A similar process, conducted with multi-ton capacity, has been commercialized recently by BASF in Germany⁷⁰. Apparently, many enantiopure amines are envisaged as targets because, being potent inhibitors of monoamine oxidase, they could be useful in the treatment of such diverse neurological disorders as Parkinson's and Alzheimer's diseases, memory loss, depression and hyperactive syndrome⁷².

Another fertile area for enzymes in organic solvents is the production of speciality polymers. For instance, by applying the enzymatic (trans)esterifications to di- or trifunctional alcohols and acids (or esters), lipase-catalysed enantioselective polycondensations in organic solvents have been achieved, leading to optically active polyesters (see ref. 73 for a review). A very different and representative application deals with peroxidase-catalysed polymerization of phenols⁷³. The resultant polyphenols may constitute an alternative to the conventional phenol-formaldehyde resins (which are worrisome owing to the recently established carcinogenicity of formaldehyde) as adhesives, laminates and photographic developers among other things. In water the oxidation of phenols with hydrogen peroxide catalysed by peroxidase results predominantly in mere dimers and trimers whose poor solubility in aqueous media is responsible for the early termination of the nascent polymer chain. In contrast, when the enzymatic polymerization is carried out in organic media, in which phenolic oligomers are freely soluble, high-molecular-mass polymers can be obtained. This process has been scaled up to a kilogram level by Enzymol International Inc. in the United States⁷³.

Intriguing opportunities seem to be offered by enzymatic polymerization in supercritical fluids. These are fluids held at temperatures and pressures above the critical point, so that the distinction between liquid and gas no longer exists. Arguably they provide an environment even more remote from water than do organic solvents. The physicochemical properties of supercritical fluids, unlike those of conventional liquids, are affected significantly by the external pressure. Because these properties can, in turn, influence the molecular mass and polydispersity of the enzymatically formed polymers, the polymeric product characteristics in a given system can be modulated by pressure⁷⁴.

Enzymes have even been shown to catalyse gas-phase reactions — that is, reactions with no condensed phase⁷⁵. Such processes presumably occur owing to the adsorption of volatile substances to a solid surface, where the enzyme is located, and subsequent enzymatic conversion there. This phenomenon seems particularly suitable for enzymatically assaying gases, such as air, containing various (for example, toxic) analytes. Analyses of this type are exemplified by the co-immobilized solid bi-enzymic system of alcohol oxidase plus peroxidase, used to detect ethanol (in human breath) or formaldehyde (for example, in the air inside a factory)⁷⁵. Either compound is first enzymatically oxidized with oxygen to give hydrogen peroxide. The latter is then taken up by peroxidase to oxidize a chromogenic substrate; the intensity of the colour thus formed is proportional to the initial concentration of the analyte.

Prospects and challenges

The ability of organic solvents, when used instead of water as reaction media, to affect and often enhance the catalytic properties of enzymes offers strategies for creating improved biocatalysts that sit alongside such techniques as site-directed mutagenesis, phage display, directed

evolution and the production of catalytic antibodies. Whereas in these protein-engineering approaches the enzyme molecule itself is modified to bring about the desired functional changes, the 'solvent engineering' described here strives to achieve such changes — including changes in catalytic activity, stability and various types of selectivity — by altering the reaction medium. It is thus potentially complementary to, and synergistic with, the means of protein engineering reviewed elsewhere in this collection of reviews.

To take full advantage of the opportunities afforded by nonaqueous enzymology, several mechanistic issues need to be elucidated. A systematic inquiry should continue into the causes of diminished enzymatic activity in nonaqueous solvents and how to prevent it¹⁵; in fact, there is no fundamental reason why enzymes could not be more active in such media than in water. Particular efforts are needed to develop a generally applicable, quantitative rationale for the solvent dependence of enzymatic selectivity^{45,46} and to ascertain the whole scope and magnitude of this promising phenomenon. The structure–function relationship of the molecular memory of enzymes in anhydrous solvents warrants further investigation. Specifically, it remains to be determined how the nature of the imprinting ligand is reflected in the modified enzymatic properties and how precise and fine-tuned the ligand-induced memory (due to cavities formed in the enzyme molecule) can be.

Although the practical utility of enzymatic catalysis in organic solvents is beyond doubt, most of the work so far has involved relatively simple, hydrolytic enzymes^{16,70}. The potential of using more complex enzymes, including those that require cofactors and especially oxidoreductases and lyases, is almost untapped. In terms of basic biochemistry, it seems that nonaqueous enzymology can provide some penetrating insights into enzyme mechanisms in general. But the research in that direction is only just beginning^{76,77}.

With the necessity of using aqueous reaction media dispelled, one can explore enzymes not only in relatively simple organic solvents and their mixtures, but also in a variety of other environments, including supercritical fluids, gases⁷⁸, eutectic mixtures⁷⁹, liquid crystals, melts and low-vapour-pressure ionic liquids. Such efforts are still in their infancy, and yet they offer intriguing opportunities for tailor-made high-performance applications. Finally, one can even look to nonaqueous whole-cell catalysis, as opposed to the isolated enzymes discussed here, for conducting complex, multistep processes. In this regard, it is encouraging that several solvent-tolerant bacterial strains have been found recently, although many basic questions remain concerning the mechanisms of solvent toxicity and possible approaches to overcoming it⁸⁰. □

1. Roberts, S. M., Turner, N. J., Willetts, A. J. & Turner, M. K. *Introduction to Biocatalysis Using Enzymes and Microorganisms* (Cambridge Univ. Press, New York, 1995).
2. Faber, K. *Biotransformations in Organic Chemistry* 4th edn (Springer, Berlin, 2000).
3. Zaks, A. & Dodds, D. R. Application of biocatalysis and biotransformations to the synthesis of pharmaceuticals. *Drug Disc. Today* **2**, 513–531 (1997).
4. Stinson, S. C. Counting on chiral drugs. *Chem. Eng. News* **76**, 83–96 (1998).
5. Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry* 3rd edn, 192 (Worth, New York, 2000).
6. Lapanje, S. *Physicochemical Aspects of Protein Denaturation* Chs 1–3 & 6 (Wiley, New York, 1978).
7. Gribenow, K. & Klibanov, A. M. On protein denaturation in aqueous-organic mixtures but not in pure organic solvents. *J. Am. Chem. Soc.* **47**, 11695–11700 (1996).
8. Kuntz, I. D. & Kauzmann, W. Hydration of proteins and polypeptides. *Adv. Protein Chem.* **28**, 239–345 (1974).
9. Rupley, J. A. & Careri, G. Protein hydration and function. *Adv. Protein Chem.* **41**, 37–172 (1991).
10. Fitzpatrick, P. A., Steinmetz, A. C. U., Ringe, D. & Klibanov, A. M. Enzyme crystal structure in a neat organic solvent. *Proc. Natl Acad. Sci. USA* **90**, 8653–8657 (1993).
11. Yennawar, N. H., Yennawar, H. P. & Farber, G. K. X-ray crystal structure of γ -chymotrypsin in hexane. *Biochemistry* **33**, 7326–7337 (1994).
12. Schmitke, J. L., Stern, L. J. & Klibanov, A. M. Crystal structure of subtilisin Carlsberg in anhydrous dioxane and its comparison with those in water and acetonitrile. *Proc. Natl Acad. Sci. USA* **94**, 4250–4255 (1997).
13. Zhu, G. *et al.* X-ray studies on two forms of bovine β -crystals in neat cyclohexane. *Biochim. Biophys. Acta* **1429**, 142–150 (1998).
14. Gao, X.-G. *et al.* Crystal structure of triosephosphate isomerase from *Trypanosoma cruzi* in hexane. *Proc. Natl Acad. Sci. USA* **96**, 10062–10067 (1999).
15. Klibanov, A. M. Enzymatic catalysis in anhydrous organic solvents. *Trends Biochem. Sci.* **14**, 141–144 (1989).
16. Koskinen, A. M. P. & Klibanov, A. M. (eds) *Enzymatic Reactions in Organic Media* (Blackie-Pergamon, London, 1996).

17. Zaks, A. & Klíbanov, A. M. Enzyme-catalyzed processes in organic solvents. *Proc. Natl Acad. Sci. USA* **82**, 3192–3196 (1985).
18. Klíbanov, A. M. Why are enzymes less active in organic solvents than in water? *Trends Biotechnol.* **15**, 97–101 (1997).
19. Schmitke, J. L., Wescott, C. R. & Klíbanov, A. M. The mechanistic dissection of the plunge in enzymatic activity upon transition from water to anhydrous solvents. *J. Am. Chem. Soc.* **118**, 3360–3365 (1996).
20. Zaks, A. & Klíbanov, A. M. Enzymatic catalysis in non-aqueous solvents *J. Biol. Chem.* **263**, 3194–3201 (1988).
21. Xu, K. & Klíbanov, A. M. pH-control of the catalytic activity of cross-linked enzyme crystals in organic solvents. *J. Am. Chem. Soc.* **118**, 9815–9819 (1996).
22. Blackwood, A. D., Curran, L. J., Moore, B. D. & Halling, P. J. Organic-phase buffers control biocatalyst activity independent of initial aqueous pH. *Biochim. Biophys. Acta* **1206**, 161–165 (1994).
23. Griebenow, K. & Klíbanov, A. M. Can conformational changes be responsible for solvent and excipient effects on the catalytic behavior of subtilisin Carlsberg in organic solvents? *Biotechnol. Bioeng.* **53**, 351–362 (1997).
24. Griebenow, K. & Klíbanov, A. M. Lyophilization-induced changes in the secondary structure of proteins. *Proc. Natl Acad. Sci. USA* **92**, 10969–10976 (1995).
25. Dabulis, K. & Klíbanov, A. M. Dramatic enhancement of enzymatic activity in organic solvents by lyoprotectants. *Biotechnol. Bioeng.* **41**, 566–571 (1993).
26. Dai, L. & Klíbanov, A. M. Striking activation of oxidative enzymes suspended in nonaqueous media. *Proc. Natl Acad. Sci. USA* **96**, 9475–9478 (1999).
27. Khmel'nitsky, Y. L., Welch, S. H., Clark, D. S. & Dordick, J. S. Salts dramatically enhance activity of enzymes suspended in organic solvents. *J. Am. Chem. Soc.* **116**, 2647–2648 (1994).
28. Russell, A. J. & Klíbanov, A. M. Inhibitor-induced enzyme activation in organic solvents. *J. Biol. Chem.* **263**, 11624–11626 (1988).
29. Broos, J., Sakodinskaya, I. K., Engbersen, J. F. J., Verboom, W. & Reinhoudt, D. N. Large activation of serine proteases by pretreatment with crown ethers. *J. Chem. Soc. Chem. Commun.* 255–256 (1995).
30. Okahata, Y. & Mori, T. Lipid-coated enzymes as efficient catalysts in organic media. *Trends Biotechnol.* **15**, 50–54 (1997).
31. Paradkar, V. M. & Dordick, J. S. Aqueous-like activity of α -chymotrypsin dissolved in nearly anhydrous organic solvents. *J. Am. Chem. Soc.* **116**, 5009–5010 (1994).
32. Margolin, A. L. Novel crystalline catalysts. *Trends Biotechnol.* **14**, 223–230 (1996).
33. Jeffrey, J. A. & Saenger, W. *Hydrogen Bonding in Biological Structures* (Springer, Berlin, 1994).
34. Affleck, R., Haynes, C. A. & Clark, D. S. Solvent dielectric effects on protein dynamics. *Proc. Natl Acad. Sci. USA* **89**, 5167–5170 (1992).
35. Burke, P. A., Griffin, R. G. & Klíbanov, A. M. Solid-state nuclear magnetic resonance investigation of solvent dependence of tyrosyl ring motion in an enzyme. *Biotechnol. Bioeng.* **42**, 87–94 (1993).
36. Zaks, A. & Klíbanov, A. M. The effect of water on enzyme action in organic media. *J. Biol. Chem.* **263**, 8017–8021 (1988).
37. Xu, Z.-F. *et al.* Transition state stabilization of subtilisins in organic media. *Biotechnol. Bioeng.* **43**, 515–520 (1994).
38. Bell, G., Halling, P. J., Moore, B. D., Partridge, J. & Rees, D. G. Biocatalyst behaviour in low-water systems. *Trends Biotechnol.* **13**, 468–473 (1995).
39. Almarsson, Ö. & Klíbanov, A. M. Remarkable activation of enzymes in nonaqueous media by denaturing organic cosolvents. *Biotechnol. Bioeng.* **49**, 87–92 (1996).
40. Ahern, T. J. & Klíbanov, A. M. Analysis of processes causing thermal inactivation of enzymes. *Methods Biochem. Anal.* **33**, 91–127 (1987).
41. Zaks, A. & Klíbanov, A. M. Enzymatic catalysis in organic media at 100 °C. *Science* **224**, 1249–1251 (1984).
42. Volkin, D. B., Staubli, A., Langer, R. & Klíbanov, A. M. Enzyme thermoinactivation in anhydrous organic solvents. *Biotechnol. Bioeng.* **37**, 843–853 (1991).
43. Garza-Ramos, G., Darszon, A., de Gómez-Puyou, M. T. & Gómez-Puyou, A. Enzyme catalysis in organic solvents with low water content at high temperatures. The ATPase of submitochondrial particles. *Biochemistry* **29**, 751–757 (1990).
44. Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York, 1999).
45. Wescott, C. R. & Klíbanov, A. M. The solvent dependence of enzyme specificity. *Biochim. Biophys. Acta* **1206**, 1–9 (1994).
46. Carrea, G., Ottolina, G. & Riva, S. Role of solvents in the control of enzyme selectivity in organic media. *Trends Biotechnol.* **13**, 63–70 (1995).
47. Zaks, A. & Klíbanov, A. M. Substrate specificity of enzymes in organic solvents vs. water is reversed. *J. Am. Chem. Soc.* **108**, 2767–2768 (1986).
48. Wescott, C. R. & Klíbanov, A. M. Solvent variation inverts substrate specificity of an enzyme. *J. Am. Chem. Soc.* **115**, 1629–1631 (1993).
49. Gaertner, H. & Puigserver, A. Kinetics and specificity of serine proteases in peptide synthesis catalyzed in organic solvents. *Eur. J. Biochem.* **181**, 207–213 (1989).
50. Ryu, K. & Dordick, J. S. How do organic solvents affect peroxidase structure and function? *Biochemistry* **31**, 2588–2598 (1992).
51. Wescott, C. R., Noritomi, H. & Klíbanov, A. M. Rational control of enzymatic enantioselectivity through solvation thermodynamics. *J. Am. Chem. Soc.* **118**, 10365–10370 (1996).
52. Ke, T., Wescott, C. R. & Klíbanov, A. M. Prediction of the solvent dependence of enzymatic prochiral selectivity by means of structure-based thermodynamic calculations. *J. Am. Chem. Soc.* **118**, 3366–3374 (1996).
53. Wu, S.-H., Chu, F.-Y. & Wang, K.-T. Reversible enantioselectivity of enzymatic reactions by media. *Bioorg. Med. Chem. Lett.* **1**, 339–342 (1991).
54. Ueki, S. *et al.* Solvent-induced inversion of enantioselectivity in lipase-catalyzed esterification of 2-phenoxypropionic acids. *Biotechnol. Lett.* **14**, 163–168 (1992).
55. Tawaki, S. & Klíbanov, A. M. Inversion of enzyme enantioselectivity by the solvent. *J. Am. Chem. Soc.* **114**, 1882–1884 (1992).
56. Rubio, E., Fernandez-Mayorales, A. & Klíbanov, A. M. Effect of the solvent on enzyme regioselectivity. *J. Am. Chem. Soc.* **113**, 695–696 (1991).
57. MacManus, D. A. & Vulfson, E. N. Reversal of regioselectivity in the enzymatic acylation of secondary hydroxyl groups mediated by organic solvents. *Enzyme Microb. Technol.* **20**, 225–228 (1997).
58. Tawaki, S. & Klíbanov, A. M. Chemoselectivity of enzymes in anhydrous media is strongly solvent dependent. *Biocatalysis* **8**, 3–19 (1993).
59. Ebert, C., Gardossi, L., Linda, P., Vesnaver, R. & Bosco, M. Influence of organic solvents on enzyme chemoselectivity and their role in enzyme-substrate interaction. *Tetrahedron* **52**, 4867–4876 (1996).
60. Klíbanov, A. M. Enzyme memory—what is remembered and why? *Nature* **374**, 596 (1995).
61. Ke, T. & Klíbanov, A. M. On enzymatic activity in organic solvents as a function of enzyme history. *Biotechnol. Bioeng.* **57**, 746–750 (1998).
62. Rich, J. O. & Dordick, J. S. Controlling subtilisin activity and selectivity in organic media by imprinting with nucleophilic substrates. *J. Am. Chem. Soc.* **119**, 3245–3252 (1997).
63. Stahl, M., Jeppson-Wistrand, U., Mansson, M.-O. & Mosbach, K. Induced stereoselectivity and substrate selectivity of bio-imprinted α -chymotrypsin in anhydrous organic media. *J. Am. Chem. Soc.* **113**, 9366–9368 (1991).
64. Mingarro, I., González-Navarro, H. & Braco, L. Trapping of different lipase conformers in water-restricted environments. *Biochemistry* **35**, 9935–9944 (1996).
65. Johnson, D. V. & Griengl, H. Biocatalytic applications of hydroxynitrile lyases. *Adv. Biochem. Eng. Biotechnol.* **63**, 31–55 (1999).
66. Ke, T. & Klíbanov, A. M. Markedly enhancing enzymatic enantioselectivity in organic solvents by forming substrate salts. *J. Am. Chem. Soc.* **121**, 3334–3340 (1999).
67. Chen, C.-S., Fujimoto, Y., Girdaukas, G. & Sih, C. J. Quantitative analyses of biochemical kinetic resolutions of enantiomers. *J. Am. Chem. Soc.* **104**, 7294–7299 (1982).
68. Bornscheuer, U. T. & Kazlauskas, R. J. *Hydrolases in Organic Synthesis—Regio- and Stereoselective Biotransformations* (Wiley, Weinheim, 1999).
69. Kirchner, G., Scollar, M. P. & Klíbanov, A. M. Resolution of racemic mixtures via lipase catalysis in organic solvents. *J. Am. Chem. Soc.* **107**, 7072–7076 (1985).
70. Carrea, G. & Riva, S. Properties and synthetic applications of enzymes in organic solvents. *Angew. Chem.* **33**, 2226–2254 (2000).
71. Kitaguchi, H., Fitzpatrick, P. A., Huber, J. E. & Klíbanov, A. M. Enzymatic resolution of racemic amines: crucial role of the solvent. *J. Am. Chem. Soc.* **111**, 3094–3095 (1989).
72. Gutman, A. L., Meyer, E., Kalerin, E., Polyak, F. & Sterling, J. Enzymatic resolution of racemic amines in a continuous reactor in organic solvents. *Biotechnol. Bioeng.* **40**, 760–767 (1992).
73. Akkara, J. A., Ayyagari, M. S. R. & Bruno, F. F. Enzymatic synthesis and modification of polymers in nonaqueous solvents. *Trends Biotechnol.* **17**, 67–73 (1999).
74. Chaudhary, A. K., Beckman, E. J. & Russell, A. J. Rational control of polymer molecular weight and dispersity during enzyme-catalyzed polyester synthesis in supercritical fluids. *J. Am. Chem. Soc.* **117**, 3728–3733 (1995).
75. Barzana, E., Karel, M. & Klíbanov, A. M. A colorimetric method for the enzymatic analysis of gases: the determination of ethanol and formaldehyde vapors using solid alcohol oxidase. *Anal. Biochem.* **182**, 109–115 (1989).
76. Marbrouk, P. A. The use of nonaqueous media to prove biochemically significant enzyme intermediates: the generation and stabilization of horseradish peroxidase compound III in neat benzene solution at room temperature. *J. Am. Chem. Soc.* **117**, 2141–2146 (1995).
77. Michels, P. C., Dordick, J. S. & Clark, D. S. Dipole formation and solvent electrostriction in subtilisin catalysis. *J. Am. Chem. Soc.* **119**, 70–76 (1997).
78. Barzana, E. Gas phase biosensors. *Adv. Biochem. Eng. Biotechnol.* **53**, 1–15 (1995).
79. Gill, I. & Vulfson, E. Enzymic catalysis in heterogeneous eutectic mixtures of substrates. *Trends Biotechnol.* **12**, 118–122 (1994).
80. De Bont, J. A. M. Solvent-tolerant bacteria in biocatalysis. *Trends Biotechnol.* **16**, 493–499 (1998).

Modular enzymes

Chaitan Khosla & Pehr B. Harbury

Departments of Chemistry, Chemical Engineering and Biochemistry, Stanford University, Stanford, California 94305, USA
(e-mail: ck@chemeng.stanford.edu)

Although modular macromolecular devices are encountered frequently in a variety of biological situations, their occurrence in biocatalysis has not been widely appreciated. Three general classes of modular biocatalysts can be identified: enzymes in which catalysis and substrate specificity are separable, multisubstrate enzymes in which binding sites for individual substrates are modular, and multienzyme systems that can catalyse programmable metabolic pathways. In the postgenomic era, the discovery of such systems can be expected to have a significant impact on the role of enzymes in synthetic and process chemistry.

A modular device is a multicomponent system in which individual components can be interchanged with functionally distinct analogues from related systems. Since the elucidation of the genetic code¹, the principle of modularity has been repeatedly uncovered in biological phenomena. Recent examples of modular mechanisms in biology have emerged from the analysis of signal transduction^{2,3} and transcriptional activation^{4,5}. In each case, the modular design of these systems has given the molecular biologist a powerful conceptual and technical base from which structure–function relationships can be probed.

In contrast to biological processes involving information transfer, metabolism is not regarded as a fertile hunting ground for modular systems. The reasons for this bias are understandable, given our deep admiration for two general properties of metabolic enzymes: their immense rate accelerations and their exquisite substrate selectivity. Because binding and catalysis are two sides of the same coin in the active site of an enzyme, the evolution of modular enzymes in which these two properties are localized in structurally distinct domains is often assumed to be incompatible with the physiological need for high reaction rates. Moreover, because most metabolic pathways involve diffusive transfer of intermediates from one enzyme to the next, maintenance of metabolic fidelity demands that an individual enzyme be able to discriminate sharply between its cognate substrate and related cellular metabolites. In this context, one might imagine that the additional requirement of modularity would present an unnecessary challenge for evolution.

Yet modular enzymes do occur in nature, although they are not common. Nonetheless, their potential utility to the chemist is enormous. Starting from a toolbox containing a reasonable number and diversity of modular enzymes, the ability to harness protein engineering to rapidly generate designer biocatalysts for any process would transform chemistry and chemical engineering. Where, then, in this postgenomic era should one look for modular enzymes? What clues can be regarded as useful indicators of the existence of modularity? Which features of enzymes can one expect to be modularized, and what mechanistic principles underlie modular operation? What are the constraints on modularity in nature's enzymes, and how might they be overcome? This review will attempt to address these questions.

Detection of modular enzymes

For an enzyme to have modular properties, it must minimally have a modular architecture. Typically this implies the existence of multiple domains (or subunits, in the case of an

oligomeric enzyme). Formally, domains are defined as stable globular fragments of proteins that may refold autonomously and carry out specific functions⁶. In practice, they are identified typically by computer algorithms that search for segments (of about 50–500 residues) with sequence similarity within a group of larger, functionally distinct polypeptides^{7–9}. There are two principal assumptions in this strategy for detecting a structurally modular enzyme. The first is that homologous domains have similar tertiary structures. Although exceptions to this rule have been observed¹⁰, this generally seems to be a reasonable assumption. More demanding is the assumption that inter-domain interactions are either absent or conserved within a family of modular enzymes. As discussed below, this represents a major limitation to the modularity of enzymes.

Whereas a domain-like architecture is suggestive of a modular enzyme, it does not guarantee functional modularity. Experimental evidence for modularity requires satisfaction of two criteria. First, distinct properties must be able to be assigned to each domain, identified by means of either proteolysis or protein engineering. Second, it should be possible to recombine these domains to generate functional chimaeras. Although very few enzyme families have been shown to satisfy both criteria for modularity, rapidly expanding sequence and structural databases continue to add to the list of potentially modular enzyme families. It should be noted that not every nominally modular

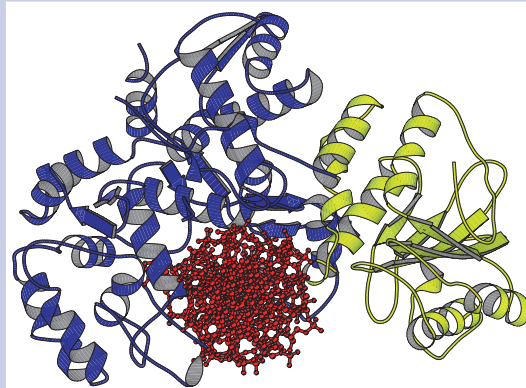


Figure 1 Separate catalytic and molecular recognition domains of the *FokI* restriction endonuclease. The carboxy-terminal DNA-cleavage domain (yellow) piggy-backs on the amino-terminal DNA-binding domain (blue). The DNA double helix is coloured red. Figure generated from the protein data bank 1FOK coordinates using MOLSCRIPT⁵³.

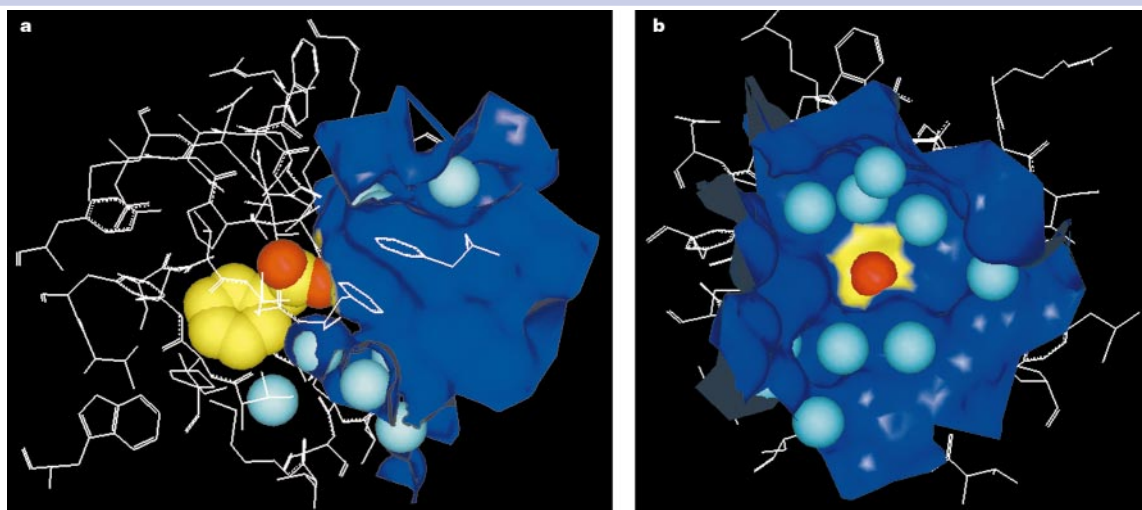


Figure 2 Recognition by penicillin G acylase of a modular chemical feature (the phenylacetyl group) in an otherwise generic substrate. Residues of penicillin acylase within 10 Å of the active site are depicted in white. The carbon and oxygen atoms of bound phenylacetic acid are coloured yellow and red respectively. A partial solvent-accessible surface of the enzyme is shown in blue. Ordered water molecules are rendered as cyan balls. **a**, Although the phenylacetyl moiety is surrounded by the enzyme, the distal oxygen atom of the acid (red) is directed into solvent. Generic amine and alcohol substrates of penicillin acylase are attached to the phenacetyl protecting group through this distal atom. **b**, A view of the same active site rotated by 90° around a vertical axis. The distal oxygen atom of phenylacetic acid (red) can be seen poking through the solvent-accessible surface. Figure generated from the 1PNL coordinates of the protein data bank using INSIGHTII (Molecular Simulations, San Diego).

biocatalyst is as perfectly modular as, for example, the machinery that catalyses ribosomal protein synthesis¹. In most cases, although the properties of individual domains or subunits can be recombined in chimaeric enzymes, the resulting chimaeras display kinetic imperfections when compared with their parents.

One can define three general categories of modular catalysts in nature: (1) enzymes in which catalysis and substrate specificity are separable; (2) multisubstrate enzymes in which the binding sites for individual substrates are modular; and (3) multienzyme systems that can catalyse programmable metabolic pathways. Using selected examples, we will illustrate the properties of enzymes belonging to each of these categories. We also discuss the opportunities for engineering enzymes in each category, as well as their potential utility in the context of applied biocatalysis. As a frame of reference, we note that the prototypical modular enzyme — the ribosomal protein biosynthetic machinery — exhibits all three types of modularity. The catalyst (the ribosome) can be readily separated from the element that dictates substrate specificity (the messenger RNA template). Moreover, the two acylated transfer RNA substrates for the peptide bond-forming

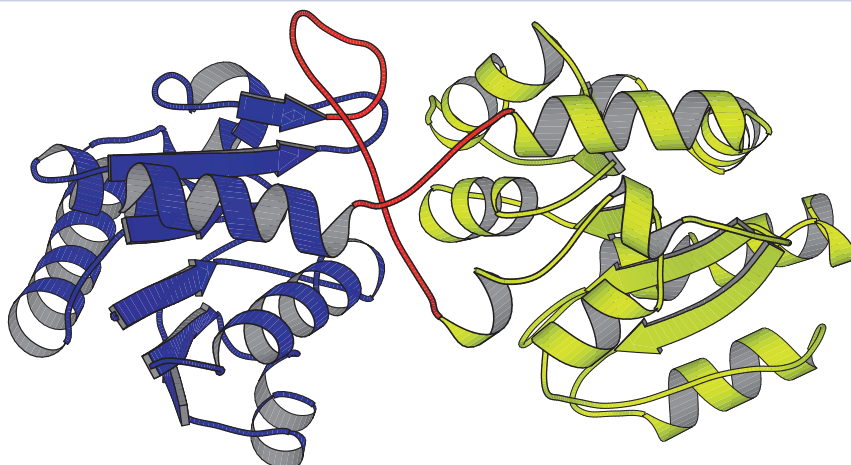
reaction bind to distinct sites in the ribosome (the A and P sites). And finally, the multistep pathway catalysed by this modular system can be reprogrammed by codon removal, addition or replacement. The same can be said about template-dependent DNA and RNA polymerases. Although such an extraordinary degree of modularity is clearly an exception rather than the norm, these modular features underlie the singular success of these systems as practically useful biocatalysts. Therefore a better understanding of the chemistry and biology of modular enzymes has important implications for biocatalysis.

Types of modular enzymes

Separation of catalysis from molecular recognition

Enzymes are prized by chemists for their regio- and stereoselectivity, yet their use is often restricted to a narrow spectrum of substrates. Paradoxically, an ideal catalyst for the chemist should be able to pinpoint a single functional group on a molecule, yet retain the ability to act on many different molecular species. If an enzyme's chemical reactivity could be separated in a modular fashion from its substrate recognition, it would be possible to create a catalyst (or family of

Figure 3 Separation of molecular-recognition features in modular multisubstrate enzymes. Crystal structure of the bi-domain MurG glycosyltransferase. The NDP-sugar and aglycone binding domains are coloured yellow and blue, respectively. The two loops that connect these two domains are coloured in red. The active-site residues lie at the interface of these two domains. For details, see ref. 36.



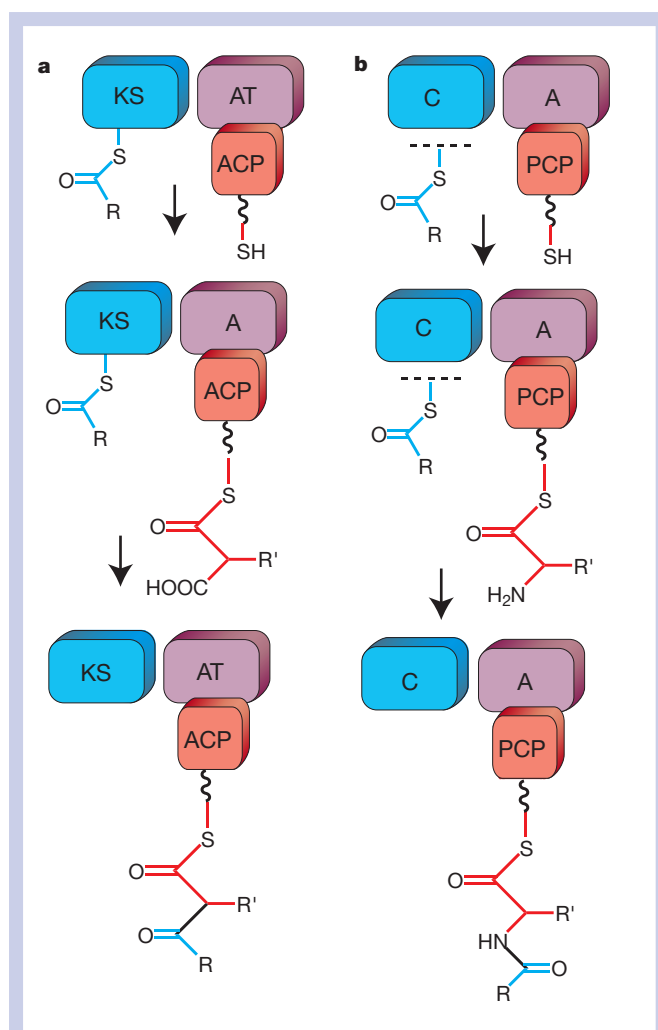


Figure 4 Separation of electrophile and nucleophile recognition in modules of polyketide synthases and non-ribosomal peptide synthetases. **a**, A polyketide synthase module is minimally comprised of three distinct domains: a condensation (KS, ketosynthase), an acyltransferase (AT), and an acyl carrier protein (ACP) domain. The electrophile (shown in blue) is recognized by and attached to the KS domain, whereas the nucleophile (shown in red) is recognized by the AT domain, which attaches this moiety to the 'swinging arm' of the ACP domain (wavy line). The C–C bond formed between the electrophile and the nucleophile is shown in black. **b**, Likewise, a non-ribosomal peptide synthetase module is also minimally comprised of three distinct domains: a condensation (C), an adenylation (A), and a peptidyl carrier protein (PCP) domain. The electrophile (shown in blue) is recognized by the C domain (although it is not believed to bind covalently to this domain, hence shown attached to the previous domain (dashed line)), whereas the nucleophile (shown in red) is recognized by the A domain, which attaches this moiety to the 'swinging arm' of the PCP domain (wavy line). The amide bond formed between the electrophile and the nucleophile is shown in black.

catalysts) that is both specific and general. This is a tall order, because substrate binding and catalytic activity are inextricably linked¹¹. Binding interactions orient substrates with respect to enzyme active-site residues, and preferentially stabilize transition states relative to ground states. Nevertheless, enzymes exploit at least three different strategies for partially uncoupling substrate binding and catalysis in a modular fashion. They include the separation of catalytic and recognition functions in distinct domains, evolution of interchangeable substrate binding pockets, and recognition of a modular chemical feature in an otherwise generic substrate.

The *FokI* restriction endonuclease exemplifies the division of labour between protein domains. It consists of two autonomously folded structures that can be separated by limited proteolysis

(Fig. 1)¹². The first domain binds a GGATG DNA-recognition motif, and the second domain catalyses DNA-strand scission nine and thirteen bases away. Fusion of the *FokI*-cleavage domain to the DNA-binding domains of unrelated transcription factors produces chimaeras with new cleavage specificities^{13,14}. However, the *FokI* system has not proved to be perfectly modular. Unexpectedly, the binding domain sequesters and inactivates the cleavage domain when the endonuclease is nonspecifically associated with DNA¹⁵. Domain-swapped enzymes that lack these inhibitory interactions exhibit high levels of nonspecific DNA cleavage.

A consortium of bacterial hydrolases that break down plant cell walls also exhibits a partial domain division between substrate recognition and chemistry. The hydrolases include xylanases and cellulases that have one or two catalytic glycosidase domains, in addition to a series of carbohydrate-binding domains (CBDs). The catalytic domains can hydrolyse soluble oligosaccharides, but they require the CBD domains to break down the crystalline sugar polymers found in plant cell walls¹⁶. The CBDs act by increasing enzyme–substrate proximity¹⁷. Fusion of cellulose- and xylan-binding domains to heterologous glycosidase domains confers the ability to metabolize crystalline substrates.

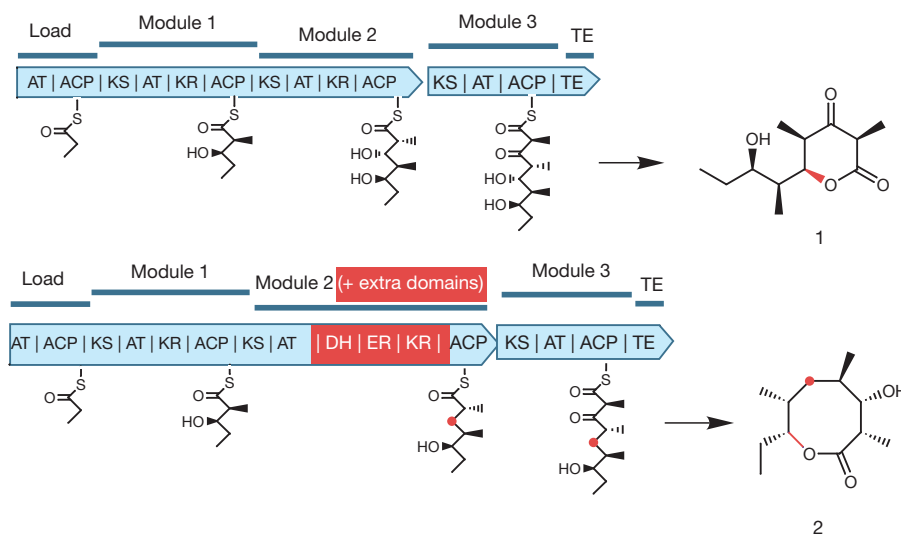
Analysis of covalent domain partnering in different genomes will probably yield other examples of enzymes with independent catalytic and binding domains¹⁸. But some mechanistic limitations of this class of enzyme should be noted. Most importantly, binding by the recognition domain does not orient the substrate for reaction. In fact, the inter-domain linker must be floppy, so that binding by the recognition domain does not prevent correct positioning of the substrate in the active site. The effective concentration of the bound substrate, held near the catalytic domain by the inter-domain linker, determines the selectivity of the enzyme. For example, if the effective concentration of a bound substrate at the active site is 10 mM, and unbound substrates are present at 1 mM, the maximum specificity exhibited by the enzyme for the bound substrate will be tenfold¹⁹.

In addition to domain-based separation, catalysis can also be uncoupled from substrate recognition by the exchange of modular binding pockets that are located in the same structural domain as the active site. The *Tetrahymena* group I ribozyme exhibits this kind of modularity²⁰. The internal guide sequence (IGS) of the ribozyme forms a duplex with the RNA substrate, positioning the substrate for cleavage. Watson–Crick base pairing between the substrate and the IGS governs substrate recognition. Base substitutions in the IGS produce ribozymes with specificity for RNA sequences bearing complementary substitutions. Likewise, the Src family of kinases provide an example of binding-pocket modularity in an enzyme with a non-polymeric substrate²¹. A single Ile→Gly substitution in the nucleotide binding pockets of Src or Fyn kinases allows the mutant proteins to efficiently use N⁶-benzyl ATP. The wild-type Src kinase discriminates against the alkylated ATP substrate by a factor of more than 400-fold.

Although multiple sequence alignments in genome data together with the growth in protein structural data will undoubtedly offer new opportunities to identify and exploit substrate binding modularity, the existence of modular substrate recognition uncoupled from catalysis is likely to be relatively rare. Chimaeric enzymes with swapped binding sites often show affinity for new substrates, but this affinity is not correlated with the ability to convert substrate to product²². Gene shuffling of enzyme families²³ may be a useful tool to shed light on the generality of binding-pocket modularity.

A third class of enzymes that modularize catalysis and molecular recognition does so by recognizing a specific chemical feature in an otherwise generic substrate. The substrate can be practically anything as long as the recognition feature is present. Examples of this type of enzyme (and the functional group recognized) include lipases (straight-chain hydrocarbons), penicillin G acylase (phenylacetyl group), butyryl cholinesterase (choline), phthalyl amidase (phthalimide) and aryl acylamidases (acetanilide)^{24–27}. Increasingly, these catalysts are being used to manipulate protecting groups on complex molecules,

Figure 5 Introduction of auxiliary catalytic domains into the module of a polyketide synthase. Formation of the six-membered lactone **1** is catalysed by a trimodular derivative of the erythromycin polyketide synthase, comprised of a loading bi-domain (Load), three consecutive modules of catalytic domains (Module 1, Module 2 and Module 3), and a terminal thioesterase (TE) domain that catalyses release of **1** from module 3 via attack of the -OH on the thioester linkage. Domains designated as KS, AT and ACP are explained in Fig. 4. Ketoreductase (KR) domains in modules 1 and 2 are responsible for generating β -OH groups on the growing chains by reducing the ketone generated in the condensation reaction. By replacing the KR domain of module 2 with a KR-DH (dehydratase)-ER (enoylreductase) tridomain from the rapamycin synthase, the β -OH (highlighted in red) is eliminated by dehydration and enoylreduction. The extra catalytic domains introduced into module 2 are therefore able to compete kinetically with the downstream module 3. Moreover, the absence of the -OH group on the tetraketide product of module 3 forces the TE domain to form the eight-membered lactone **2**. For details, see ref. 54.



which are sensitive to the harsh chemical reagents of traditional organic chemistry, or to produce chiral molecules^{28–30}. For example, the *Serratia marcescens* lipase is used to manufacture a synthetic intermediate of the calcium antagonist diltiazem on the scale of 50 tons yr⁻¹.

Enzymes with the unusual ability to recognize an isolated chemical module have so far been discovered empirically. The recent crystal structures of penicillin G acylase (Fig. 2) and of several lipases offer mechanistic insights into their mode of action. Both structures reveal shallow active sites, with variable regions of the substrate directed outwards into solvent^{31–33}. Large portions of the substrate are not in intimate contact with the enzyme. Where might one look for additional examples of this class of catalyst? Bacterial lipases, phthalyl amidase and penicillin acylase are thought to function as carbon-source scavengers for the microorganisms that produce them. Aryl acylamidases in plants, rhizobacteria and soil microorganisms break down diverse acetanilide-derived pesticides (such as propanil), thereby conferring resistance. Butyryl cholinesterase exists in the liver and plasma of humans. It scavenges general choline esters and detoxifies a large number of drugs, including cocaine. The pattern of natural activities indicates that enzymes involved in nutrient recruitment and chemical warfare may have evolved under selective pressure to tolerate diverse substrates. Consequently, proteins involved in these processes might be expected to provide a rich source of undiscovered catalysts for enzymatic biotransformation.

Separation of molecular-recognition features in multisubstrate enzymes

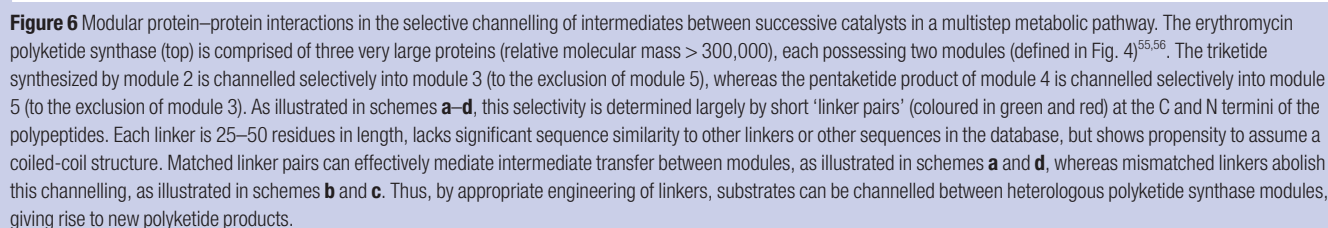
Many enzymes catalyse reactions involving two (or more) substrates. Broadly speaking, these enzymes fall into two categories. Reactions in which all substrates bind to the enzyme before the first product is formed are called sequential, whereas those in which one or more products are released are called ping-pong³⁴. The active sites of sequential enzymes can bind substrates in a random or a defined order. In ping-pong enzymes or in ordered sequential enzymes, the binding pockets for both substrates often overlap or interact with each other. But in random-binding, bi-substrate enzymes, the binding pockets are generally separated; indeed, in some cases they may even lie within architecturally distinct domains of the polypeptide backbone. A vivid example of such modular enzymes is methionine synthase, whose substrates — homocysteine, methyltetrahydrofolate, cobalamin and adenosylmethionine — bind to distinct regions of the protein³⁵. Modularity of molecular-recognition

features of multisubstrate enzymes represents a fertile starting point for protein engineering.

Sequence and structural analysis also indicates that many families of evolutionarily related bi-substrate enzymes have modular molecular-recognition features. For example, enzymes that use nucleoside diphosphate (NDP)-sugars, such as glycosyltransferases³⁶ and dehydrogenases³⁷, seem to have this property. NDP-glycosyltransferases in particular belong to a large family of enzymes that are known to possess relaxed specificity for both the sugar and the aglycones^{38,39}. Recently the crystal structure of a prototype of this family has revealed a two-lobed architecture (Fig. 3)³⁶. One domain binds to the NDP-sugar whereas the other domain binds to the aglycone. The two domains are separated by extended loops, and the active-site residues lie at the interface of the two domains. Sequence comparisons indicate that most members of this enzyme superfamily retain both the bi-domain architecture and the location of active-site residues⁴⁰. Whereas functional evidence for modularity is lacking, it is plausible that the molecular-recognition features of these two domains have not only evolved independently but also been exchanged frequently without destroying the active-site geometry. If so, then the relaxed specificity of individual domains together with the biosynthetic importance of this enzyme family makes it a particularly attractive target for domain shuffling.

The individual modules of polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) represent another strategy that nature seems to have exploited for modularizing molecular recognition features in protein catalysts. Here, an electrophile and a nucleophile are attached covalently to two distinct domains (or subunits) of a multidomain (or multisubunit) enzyme (Fig. 4)⁴¹. Selectivity for the electrophile resides in the domain that catalyses bond formation between the two substrates (the condensing enzyme), whereas nucleophile selectivity is controlled by a transfer domain, which attaches the nucleophile onto the pantetheine arm of a carrier domain. Although direct structural evidence for the modularity of these three domains is lacking, numerous studies have shown that catalytic bond formation between the two substrates can be reconstituted by recombination of heterologous condensing, transfer and carrier domains (for reviews, see refs 42 and 43).

In contrast to the relative rarity of enzymes that modularize catalysis and molecular recognition, multisubstrate enzymes with modular recognition features are probably more common in nature.



What are the limits to the modularity of the substrate-recognition features of catalysts such as glycosyltransferases, PKS modules or NRPS modules? Analysis of individual PKS and NRPS modules has demonstrated that the kinetics of catalytic bond formation is influenced by the selection of both the electrophile and the nucleophile substrates^{44,45}. Although these ‘imperfections’ in modularity do not seem to present qualitative barriers to the predictive design and biosynthetic engineering of new ‘unnatural’ natural products, they do affect the productivity of hybrid multifunctional catalysts in fermentation processes⁴⁶. A combination of random mutagenesis and structure-based approaches may be useful in ameliorating these bottlenecks.

Systems such as PKs and NRPSs illustrate yet another principle for modular catalysis that, if generalized, would have significant implications for the development of one-pot biocatalytic processes. Instead of relying strictly on diffusion, they have evolved a highly modular strategy to channel intermediates from one active site to the next. Channelling can be defined as the direct transfer of an intermediate between consec-

Multienzyme assemblies such as α -keto-dehydrogenases, fatty acid synthases, PKSs and NRPSs have 'swinging arms' (lipoamide in the case of dehydrogenases⁵¹, and phosphopantetheine in the cases of fatty acid synthases, PKSs and NRPSs⁴¹). These swinging arms are flexible, long (~10–15 Å) tethers that channel covalently bound intermediates between successive active sites (see Fig. 4). Two examples illustrate possible mechanisms by which a swinging arm can be combined with other structural features to make a metabolic pathway modular. First, individual PKS modules have been altered to expand their repertoire of catalytic functions. Gain-of-function mutagenesis involves grafting auxiliary catalytic domains into a core PKS module, and can be used to introduce new chemistry into the reaction sequence catalysed by the module (Fig. 5). Not only does this highlight the extraordinary structural plasticity of a module, but it also suggests that the presence of a swinging arm provides an effective mechanism for newly grafted domains to compete for potential substrates before they are transferred from one module to the next. Second, in combination with selective protein–protein interactions, swinging arm-mediated chemistry can facilitate transfer of natural and unnatural intermediates between mod-

ules by attenuating the role of protein–substrate interactions (Fig. 6; S. Y. Tsui, D. E. Cane and C.K., unpublished results). Here metabolism seems to have borrowed a chapter directly from signal-transduction mechanisms by incorporating selectivity into matched pairs of short ‘linkers’ — as exemplified by the case of Fos–Jun interactions, where modular coiled-coil segments stabilize the heterodimer preferentially over either homodimer⁵².

What limits more widespread exploitation of modular channelling mechanisms in multistep bioconversion processes? First, these mechanisms seem to be restricted to systems that rely extensively on covalent catalysis, presumably because of the need for a swinging arm. Second, even where selective linkers can be engineered to direct intermediates between designated catalytic modules, in themselves they can at most provide the advantage of intramolecularity. For some catalytic reactions, approximation can provide a huge rate enhancement; in other cases it is only of relatively modest value. Finally, not all protein–protein interactions are modular. For example, in a PKS module, if interactions between the donor carrier domain and the acceptor condensing enzyme (ACP₂ and KS₃ in Fig. 6) are important to chain transfer, these properties will vary with module ultrastructure and will be difficult to categorize universally.

Conclusions

The brief history of molecular and cellular biology has demonstrated repeatedly that modularity in biological macromolecules can be exploited by both evolution and engineering. In many such enzymes, the structural and mechanistic basis for modularity is only now being elucidated. Yet notwithstanding our rudimentary knowledge about these enzymes, their utility in practical biocatalysis has been well established, and their attractiveness as targets for protein engineering is becoming apparent. As our understanding of these remarkable catalysts advances, and as the tools of molecular biology and knowledge-based protein design improve, one can expect to see more such engineered enzymes making the transition from the proof-of-principle stage to industrially useful biocatalysts with respectable space-time yields. At the same time, as sequence and structural databases continue to grow, new families of modular enzymes will surely emerge to expand the repertoire of chemistry that is accessible to modular biocatalysts. □

1. Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
2. Cohen, G. B., Sen, R. & Baltimore, D. Modular binding proteins in signal transduction proteins. *Cell* **80**, 237–248 (1995).
3. Sudol, M. From Src homology domains to other signaling modules: proposal of the protein recognition code. *Oncogene* **17**, 1469–1474 (1998).
4. Frankel, A. D. & Kim, P. S. Modular structure of transcription factors: implications for gene regulation. *Cell* **65**, 717–719 (1991).
5. Ptashne, M. & Gann, A. Imposing specificity by localization: mechanism and evolvability. *Curr. Biol.* **8**, R897–R904 (1998).
6. Rashin, A. A. Location of domains in globular proteins. *Nature* **291**, 85–87 (1981).
7. Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372**, 631–634 (1994).
8. Henikoff, S. *et al.* Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609–614 (1997).
9. Riley, M. & Labedan, B. Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. *J. Mol. Biol.* **268**, 857–868 (1997).
10. Dalal, S., Balasubramanian, S. & Regan, L. Protein alchemy: changing β -sheet into α -helix. *Nature Struct. Biol.* **4**, 548–552 (1997).
11. Jencks, W. P. *Catalysis in Chemistry and Enzymology* 615–806 (Dover, New York, 1987).
12. Li, L., Wu, L. P. & Chandrasegaran, S. Functional domains in Fok-I restriction endonuclease. *Proc. Natl Acad. Sci. USA* **89**, 4275–4279 (1992).
13. Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes—zinc finger fusions to Fok I cleavage domain. *Proc. Natl Acad. Sci. USA* **93**, 1156–1160 (1996).
14. Huang, B. H., Schaeffer, C. J., Li, Q. H. & Tsai, M. D. Splase—a new class I zinc-finger restriction endonuclease with specificity for Sp1 binding sites. *J. Protein Chem.* **15**, 481–489 (1996).
15. Aggarwal, A. K. & Wah, D. A. Novel site-specific DNA endonucleases. *Curr. Opin. Struct. Biol.* **8**, 19–25 (1998).
16. Fernandes, A. C. *et al.* Homologous xylanases from *Clostridium thermocellum*: evidence for bi-functional activity, synergism between xylanase catalytic modules and the presence of xylan-binding domains in enzyme complexes. *Biochem. J.* **342**, 105–110 (1999).
17. Bolam, D. N. *et al.* *Pseudomonas* cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. *Biochem. J.* **331**, 775–781 (1998).
18. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753 (1999).

19. Jencks, W. P. On the attribution and additivity of binding energies. *Proc. Natl Acad. Sci. USA* **78**, 4046–4050 (1981).
20. Murphy, F. L. & Cech, T. R. Alteration of substrate specificity for the endoribonucleolytic cleavage of RNA by the *Tetrahymena* ribozyme. *Proc. Natl Acad. Sci. USA* **86**, 9218–9222 (1989).
21. Liu, Y., Shah, K., Yang, F., Witucki, L. & Shokat, K. M. Engineering Src family protein kinases with unnatural nucleotide specificity. *Chem. Biol.* **5**, 91–101 (1998).
22. Hedstrom, L., Farrjones, S., Kettner, C. A. & Rutter, W. J. Converting trypsin to chymotrypsin—ground-state binding does not determine substrate specificity. *Biochemistry* **33**, 8764–8769 (1994).
23. Stemmer, W. P. C. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389–391 (1994).
24. Jaeger, K. E., Dijkstra, B. W. & Reetz, M. T. Bacterial biocatalysts: molecular biology, three-dimensional structures, and biotechnological applications of lipases. *Annu. Rev. Microbiol.* **53**, 327–351 (1999).
25. Pathak, T. & Waldmann, H. Enzymes and protecting group chemistry. *Curr. Opin. Chem. Biol.* **2**, 112–120 (1998).
26. Costello, C. A., Kreuzman, A. J. & Zmijewski, M. J. Selective deprotection of phthalyl protected amines. *Tetrahedron Lett.* **37**, 7469–7472 (1996).
27. Villarreal, D. T., Turco, R. F. & Konopka, A. A structure-activity study with aryl acylamidases. *Appl. Env. Microbiol.* **60**, 3939–3944 (1994).
28. Roberts, S. M. Preparative biotransformations: the employment of enzymes and whole-cells in synthetic organic chemistry. *J. Chem. Soc. Perkin Trans. 1*, 157–169 (1998).
29. Roberts, S. M. Preparative biotransformations. *J. Chem. Soc. Perkin Trans. 1*, 1–21 (1999).
30. Roberts, S. M. Preparative biotransformations. *J. Chem. Soc. Perkin Trans. 1*, 611–633 (2000).
31. Duggleby, H. J. *et al.* Penicillin acylase has a single amino-acid catalytic centre. *Nature* **373**, 264–268 (1995).
32. Done, S. H., Brannigan, J. A., Moody, P. C. E. & Hubbard, R. E. Ligand-induced conformational change in penicillin acylase. *J. Mol. Biol.* **284**, 463–475 (1998).
33. Lang, D. A., Mannesse, M. L. M., DeHaas, G. H., Verheij, H. M. & Dijkstra, B. W. Structural basis of the chiral selectivity of *Pseudomonas cepacia* lipase. *Eur. J. Biochem.* **254**, 333–340 (1998).
34. Fersht, A. *Enzyme Structure and Mechanism* (Freeman, 1985).
35. Goulding, C. W., Postigo, D. & Matthews, R. G. Cobalamin-dependent methionine synthase is a modular protein with distinct regions for binding homocysteine, methyltetrahydrofolate, cobalamin, and adenosylmethionine. *Biochemistry* **36**, 8082–8091 (1997).
36. Ha, S., Walker, D., Shi, Y. & Walker, S. The 1.9 Å crystal structure of *Escherichia coli* MurG, a membrane-associated glycosyltransferase involved in peptidoglycan biosynthesis. *Protein Sci.* **9**, 1045–1052 (2000).
37. Campbell, R. E., Mosimann, S. C., van de Rijn, I., Tanner, M. E. & Strynadka, N. C. The first structure of UDP-glucose dehydrogenase reveals the catalytic residues necessary for the two-fold oxidation. *Biochemistry* **39**, 7012–7023 (2000).
38. Borisova, S. A., Zhao, L., Sherman, D. H. & Liu, H. W. Biosynthesis of desosamine: construction of a new macrolide carrying a genetically designed sugar moiety. *Org. Lett.* **1**, 133–136 (1999).
39. Olano, C., Lomovskaya, N., Fontein, L., Roll, J. T. & Hutchinson, C. R. A two-plasmid system for the glycosylation of polyketide antibiotics: bioconversion of epsilon-rhodomyconone to rhodomycon D. *Chem. Biol.* **6**, 845–855 (1999).
40. Kapitonov, D. & Yu, R. K. Conserved domains of glycosyltransferases. *Glycobiology* **9**, 961–978 (1999).
41. Cane, D. E., Walsh, C. T. & Khosla, C. Harnessing the biosynthetic code. Combinations, permutations, mutations. *Science* **282**, 63–68 (1998).
42. Khosla, C., Gokhale, R., Jacobsen, J. R. & Cane, D. E. Tolerance and specificity of polyketide synthases. *Annu. Rev. Biochem.* **68**, 219–253 (1999).
43. Mootz, H. D. & Marahiel, M. A. Design and application of multimodular peptide synthetases. *Curr. Opin. Biotechnol.* **10**, 341–348 (1999).
44. Belshaw, P. J., Walsh, C. T. & Stachelhaus, T. Aminoacyl-CoAs as probes of condensation domain selectivity in nonribosomal peptide synthesis. *Science* **284**, 486–489 (1999).
45. Wu, N., Kudo, F., Cane, D. E. & Khosla, C. Analysis of the molecular recognition features of individual modules derived from the erythromycin polyketide synthase. *J. Am. Chem. Soc.* **122**, 4847–4852 (2000).
46. McDaniel, R. *et al.* Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel “unnatural” natural products. *Proc. Natl Acad. Sci. USA* **96**, 1846–1851 (1999).
47. Geck, M. K. & Kirsch, J. F. A novel, definitive test for substrate channeling illustrated with the aspartate aminotransferase/malate dehydrogenase system. *Biochemistry* **38**, 8032–8037 (1999).
48. Hyde, C. C., Ahmed, S. A., Padlan, E. A., Miles, E. W. & Davies, D. R. Three-dimensional structure of the tryptophan synthase alpha 2 beta 2 multienzyme complex from *Salmonella typhimurium*. *J. Biol. Chem.* **263**, 17857–17871 (1988).
49. Krahn, J. M. *et al.* Coupled formation of an amidotransferase interdomain ammonia channel and a phosphoribosyltransferase active site. *Biochemistry* **36**, 11061–11068 (1997).
50. Thoden, J. B., Holden, H. M., Wesenberg, G., Raushel, F. M. & Rayment, I. Structure of carbamoyl phosphate synthetase: a journey of 96 Å from substrate to product. *Biochemistry* **36**, 6305–6316 (1997).
51. Perham, R. N. & Reche, P. A. Swinging arms in multifunctional enzymes and the specificity of post-translational modification. *Biochem. Soc. Trans.* **26**, 299–303 (1998).
52. O’Shea, E. K., Rutkowski, R., Stafford, W. F. III & Kim, P. S. Preferential heterodimer formation by isolated leucine zippers from Fos and Jun. *Science* **245**, 646–648 (1989).
53. Kraulis, P. J. Molscript—a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
54. Kao, C. M. *et al.* Gain of function mutagenesis of a modular polyketide synthase II. Engineered biosynthesis of an eight-membered ring tetraketide lactone. *J. Am. Chem. Soc.* **119**, 11339–11340 (1997).
55. Cortes, J., Haydock, S. F., Roberts, G. A., Bevit, D. J. & Leadlay, P. F. An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of *Saccharopolyspora erythraea*. *Nature* **348**, 176–178 (1990).
56. Donadio, S., Staver, M. J., McAlpine, J. B., Swanson, S. J. & Katz, L. Modular organization of genes required for complex polyketide biosynthesis. *Science* **252**, 675–679 (1991).

Acknowledgements

Research on modular enzymes in C.K.’s laboratory is supported by grants from the National Science Foundation and the National Institutes of Health. P.B.H. is a Terman Fellow, a Searle Scholar and a Burroughs–Wellcome Young Investigator in the Pharmacological Sciences. We thank S. Walker for helpful discussions regarding glycosyltransferases.

Combinatorial and computational challenges for biocatalyst design

Frances H. Arnold

Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, USA

Nature provides a fantastic array of catalysts extremely well suited to supporting life, but usually not so well suited for technology. Whether biocatalysis will have a significant technological impact depends on our finding robust routes for tailoring nature's catalysts or redesigning them anew. Laboratory evolution methods are now used widely to fine-tune the selectivity and activity of enzymes. The current rapid development of these combinatorial methods promises solutions to more complex problems, including the creation of new biosynthetic pathways. Computational methods are also developing quickly. The marriage of these approaches will allow us to generate the efficient, effective catalysts needed by the pharmaceutical, food and chemicals industries and should open up new opportunities for producing energy and chemicals from renewable resources.

Biological systems are masterful chemists — a fact long appreciated by those who study how living things build complex molecules and systems from simple compounds. Enzymes catalyse the interconversion of a vast number of molecular structures, achieving tasks that range from the fixation of nitrogen to the synthesis of large and intricately structured molecules that ward off predators or attract mates. Such catalysts are models of energy-efficient, environmentally benign chemical agents, as virtually all do their work under mild conditions — in water, at room temperature and atmospheric pressure — and generate few waste products.

In view of increasing environmental and economic pressure to use renewable sources for energy and chemical feedstocks in industry, biocatalysts look like potentially attractive technological tools. But enzymes have evolved to contribute to the survival and reproduction of the organisms that make them; that they might also be useful in laundry detergents or to synthesize a new drug is simply serendipitous. In fact, attempts to use enzymes or whole organisms in applied chemical processes or products reveal some severe disadvantages of biocatalysts. Some of them turn off when a little product accumulates. This feature, so useful in regulating the flow of metabolites inside a cell, quickly derails implementation of a biocatalytic process to make that product. The process engineer is also unlikely to favour a delicate catalyst that must be replaced every few hours or must be coddled to keep it going. And of course nature does not conveniently provide a catalyst for any transformation we wish to conduct.

For many years the identification of new biocatalysts depended on labour-intensive screening of microbial cultures for the desired activities. Almost all the biocatalysts in use today came from the small fraction of organisms that can be grown under controlled conditions, the 'microbial weeds'. (By most counts less than 1% of all microorganisms can be cultured.) Some of these organisms live in harsh environments and their catalysts exhibit remarkable and useful properties, including the ability to function under extreme conditions of temperature, salt or pH¹. Other organisms have potentially useful new catalysts or enzyme pathways that allow them to produce valuable, biologically active compounds². These catalysts could be recruited to make the natural products in more tractable organisms such as *Escherichia coli*.

Efforts to comb natural biodiversity for useful activities have been greatly facilitated by high-throughput screening technologies and by new methods for collecting genes from the environment and expressing them in recombinant organisms³⁻⁴. These processes allow faster access to useful catalytic activities from organisms that cannot be cultured. But natural diversity cannot address all practical biocatalytic problems. Screening larger libraries of DNA or microbes may not even be the fastest or most efficient route to obtaining a good catalyst. Some problems can be solved by the right method of implementation — immobilization or crystallization can stabilize weak protein structures, for example. But many problems are best attacked by engineering the catalyst itself, whether it be a single enzyme, multiple enzymes or even a whole cell.

A revolution in biological design possibilities was unleashed by the advent of recombinant DNA technology, with which one can manipulate DNA sequences in a highly specific fashion and express their protein products in a variety of organisms, from animals to bacteria. This provides a means to redesign nature's catalysts at the molecular level according to detailed specifications, and to produce them in large quantities in fast-growing microorganisms. In this review I consider the ways in which biotechnological methods permit the restructuring of enzymes to adapt their functions for applied ends. Broadly speaking, one can identify two philosophies: either existing biocatalysts can be fine-tuned by rational redesign, or combinatorial techniques can be used to search for useful functionality in libraries generated at random and improved by suitable selection methods.

Our ability to manipulate the structures and functions of biological molecules and even whole organisms at will carries the prospect of applications previously not considered in the realm of biocatalysis, such as very large scale chemical production. In the not-too-distant future we can expect custom-made enzymes for gene therapies^{5,6}, new reagents for basic science and clinical diagnostics^{7,8}, and even new designs of the cellular machinery for making proteins *in vivo*⁹.

Rational redesign of natural biocatalysts

To take full advantage of recombinant DNA technology for making new enzymes, we need to know the connection between protein sequence and function. In other words, redesigning nature's catalysts rationally — that is, by

specifying the sequence — usually requires detailed understanding of structures and mechanisms. This information is unavailable for the vast majority of enzymes. Even if the target enzyme is well characterized, the molecular basis for the desired function may not be. With hundreds and even thousands of atoms that interact weakly with each other in an ensemble of closely related and interconverting folded conformations, the complex and finely tuned enzyme fades easily in the clumsy hands of the protein engineer.

Despite these challenges, biological design is now going through its most exciting period since the introduction of recombinant DNA methods and the invention of site-directed mutagenesis over two decades ago. One factor contributing to this capability is the exponential growth of the databases of protein structures and sequences. By arranging proteins in family trees we have learnt that those proteins existing today evolved from a much smaller set of ancient molecules. Shuffling of segments, fusions with other proteins and accumulation of random mutations all contributed to their diversification. Sometimes the sequences remained sufficiently conserved during this process that we can compare the sequence of a new biocatalyst identified in a screening programme to the thousands that have been deposited in databases and identify related proteins whose functions, and maybe even structures, are already known. With this information we can make inferences about the new catalyst's structure and activity.

There is now ample evidence that new enzymes evolved in nature by relatively minor modification of active-site structures^{10,11}. Thus sequence and structure information can sometimes be used to good effect in transferring activities from one enzyme to another related one. Shanklin and co-workers have exploited this notion to re-engineer membrane-bound di-iron enzymes that exhibit distinct hydroxylase, epoxidase, acetylenase and conjugase activities during fatty-acid biosynthesis in plants^{12,13}. It is believed that all these activities arose from a common progenitor enzyme through modification of the active site to allow direction of a common free-radical reaction intermediate into the different end products¹⁴ (Fig. 1). Comparing the sequences of five related oleate desaturases with those of two hydroxylases, Shanklin and co-workers identified seven positions that were strictly conserved within the five desaturases but which differed from the equivalent positions in the two hydroxylases¹². Reciprocal amino-acid changes between one of the desaturases and one of the hydroxylases at the seven sites yielded pronounced shifts in the ratio of desaturase to hydroxylase activity. Four amino-acid substitutions were sufficient to convert the desaturase to a hydroxylase, and as few as six substitutions turn the hydroxylase into a desaturase.

Another example comes from work on a dehalogenase from the 2-enoyl-coenzyme A (CoA) hydratase/isomerase superfamily¹¹. Comparison of the sequences and structures of various family members revealed that their active sites can be viewed as derivatives of a single active-site structure that provides CoA binding, an oxyanion pocket and a chamber containing stations at which substrate binding and catalytic groups have been strategically positioned. Experiments with site-directed mutagenesis show that chemical diversification can be achieved through placement of one or more polar residues along the stations: grafting eight amino-acid substitutions from crotonyl hydratase conferred on its relative, 4-chlorobenzoyl-CoA dehalogenase, the new ability to catalyse the hydration of crotonyl-CoA.

This comparative approach can be useful for identifying amino acids that control particular enzyme behaviours and demonstrating mechanisms for the diversification of catalytic functions in nature. Catalytic plasticity has clearly contributed to the evolution of chemical diversity. But such comparisons are likely to be of limited use for designing new biocatalysts, because the changes produced by altering the identified amino acids often do not extend beyond the range encoded in the parental genes¹³. One of the main goals of biocatalyst engineering is to endow them with new features that are not found in the natural sequences because they confer no evolutionary advantage.

A further problem is that the amino acids care about their context — their neighbours influence their contributions to an enzyme's

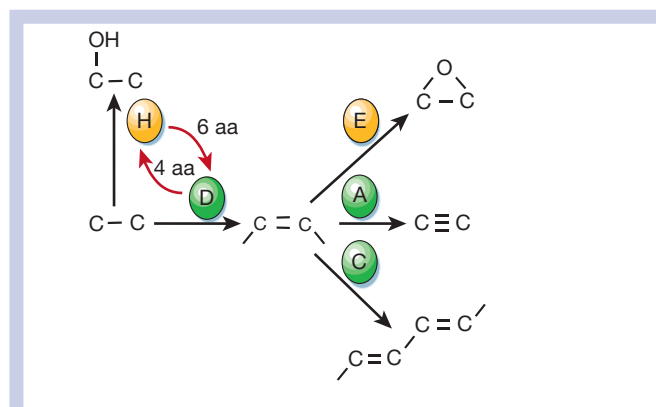


Figure 1 Catalytic plasticity in a family of fatty-acid synthesis enzymes. A family of closely related (>50% amino-acid identity) Fad2-desaturase-like lipid-modification enzymes can mediate a range of functional outcomes. Enzymes: D, oleate desaturase; H, oleate hydroxylase; E, linoleate epoxidase; A, linoleate acetylenase; C, linoleate conjugase. Yellow enzymes denote oxygen transfer; green enzymes denote hydrogen abstraction. Red arrows indicate the number of amino-acid (aa) substitutions shown to substantially affect reaction outcome^{13,14}.

activity. Also, many important biocatalyst properties are not localized in a small number of catalytic residues, but reflect contributions from many residues distributed over large parts of the protein. Even when large functional changes can be obtained with a few amino-acid substitutions, it will often be difficult or impossible to discern the specific mutations responsible. Most sequence changes accumulated during evolution have little or no effect on the property of interest, and their presence makes it difficult to pick out the key positions¹⁵. A good example is stability. Hundreds of amino acids using many types of interactions can contribute to the stability of a protein, and useful design rules for stabilization have not yet been extracted from sequence comparisons¹⁶. The factors determining stability have, however, been good targets for powerful computational methods of protein modelling that can handle the large numbers of competing interactions^{17,18}.

Nearly all engineered enzymes that are used today came out of structure-based protein-engineering efforts of the 1980s. The successes have been notable, but the results were costly and came far too slowly. Although some properties, notably enzyme specificity, respond relatively well to structure-based design and site-directed mutagenesis¹⁹, this approach is often cumbersome and unsatisfactory for engineering industrial biocatalysts which must meet a long list of performance specifications and for which the windows of opportunity are all too brief. In the pharmaceuticals industry, new catalysts must be selected and implemented often within a few months. Predictive capabilities are still rudimentary for catalysis, and even when successful, the desired changes in activity and specificity often come at the cost of other, equally important properties, such as stability or expression level.

Breeding a better catalyst

Another key factor contributing to expanding biological design capabilities is the development of 'evolutionary' protein design methods²⁰ using random mutagenesis, gene recombination and high-throughput screening.

Unlike natural evolution, laboratory evolution is directed — more like breeding^{21,22}. A 'generation' of molecules can be bred in a few days, with large numbers of progeny subject to selective pressures not encountered in nature. Because the molecules are produced in recombinant cells and are decoupled from their biological functions, they can be bred for non-natural but useful properties, including the ability to carry out reactions on substrates not encountered in nature, or to function under highly unusual conditions. Because molecules can be bred for multiple traits simultaneously by changing the

conditions of the screen or selection, this approach is particularly attractive for engineering industrial biocatalysts.

Although there are many ways to evolve a biocatalyst in the laboratory, they all involve two main steps: making a set of mutant biocatalysts and searching that set for mutants with the desired properties. The process can be iterative, so that large changes in function are obtained by accumulating small changes over many generations.

Sequential rounds of random mutagenesis carried out on ever-improved mutants is a simple and highly effective strategy that has been applied successfully to a number of catalyst design problems. Particularly relevant to biocatalysis — and particularly difficult to manipulate using structure-based design — is enzyme enantioselectivity²³. Subtle changes in enzyme structure and even changes in reaction conditions can influence enantioselectivity, but these effects are almost impossible to predict. Enantioselectivity can, however, be tuned by laboratory evolution. Starting from a naturally occurring lipase with almost no selectivity for the hydrolysis of racemic 2-methyldecanoic ester, Liebeton *et al.*²⁴ evolved an enzyme that catalysed the reaction at more than 90% enantiomeric excess using several rounds of mutagenesis and screening. In another recent study, three generations of mutagenesis and screening actually inverted the enantioselectivity of a hydantoinase to prefer L- over D-5-(2-methylthioethyl) hydantoin and increased its activity fivefold²⁵. Degussa AG is currently evaluating a whole-cell catalyst incorporating the evolved enzyme for commercial production of enantiopure L-methionine.

Laboratory evolution has also been effective in altering other key biocatalyst properties, including stability, function in non-natural environments (such as organic solvents; see accompanying review by Klivanov, pages 241–246), product inhibition, expression in a recombinant host and substrate specificity^{21,23,26}. A particularly impressive example is the evolution of an aspartate transaminase to have 2.6×10^6 -fold higher activity towards the non-native substrate valine^{27,28}. The crystal structure of the evolved enzyme shows how the active site was remodelled through the cumulative effects of mutations distributed over much of the enzyme structure²⁷. Yet only one of the 17 mutated residues contacts the substrate, and none contact the pyridoxal 5'-phosphate cofactor. This study illustrates well how complex the solutions to enzyme design problems can be, a point echoed in structural analyses of other laboratory-evolved enzymes²⁹. In the right places, amino acids serve as 'molecular shims'¹³ to tune

substrate and reaction specificity; beneficial amino-acid substitutions easily identified by random mutagenesis and screening may have minute structural consequences, beyond the resolution of structural analysis and certainly beyond our ability to predict.

Accumulating point mutations is an effective fine-tuning mechanism, but nature also uses other means to create new molecular diversity on which evolution can act. One of those is recombination. Recent studies show that recombination is an extremely useful operation for laboratory evolution. So-called DNA shuffling methods³⁰ pioneered by Stemmer create hybrid gene libraries by homologous recombination of related parent genes (ref. 31 and Fig. 2). This 'molecular sex' creates new genes that code for proteins with sequence information from any or all parents. Genes from multiple parents and even from different species can be shuffled in a single step, operations that are forbidden in nature but may be very useful for rapid adaptation. DNA shuffling is used widely to generate highly improved biocatalysts, as well as ones with features not present in the parent enzymes and not known to occur in nature^{21,22}.

This molecular breeding concept extends nicely to more complex problems involving many interacting genes. A good example is creating new, multienzyme pathways for making chemicals. Microorganisms, plants and animals produce a wide range of compounds that could function as new drugs, dyes, fragrances, flavourings and cosmetics. But many are found only in trace quantities in their natural sources and are difficult or impossible to synthesize chemically. An important goal for biocatalysis is to produce these compounds in fast-growing organisms suitable for large-scale production.

Genes encoding the enzymes that catalyse the series of chemical reactions necessary to make a particular compound can be transferred to more amenable host organisms, conferring on them the new ability to synthesize the desired product^{32,33}. Molecular breeding can optimize the engineered pathways, and it can also create new pathways, capable of synthesizing novel compounds.

Schmidt-Dannert *et al.*³⁴ have evolved the pathways that synthesize carotenoid pigments. Using a small set of bacterial genes that produce β -carotenes, they were able to exploit the remarkable plasticity of carotenoid biosynthetic pathways to generate pathways for a number of related carotenoids and precursors (Fig. 3). The two genes from *Erwinia* sp. that produce phytoene were engineered into

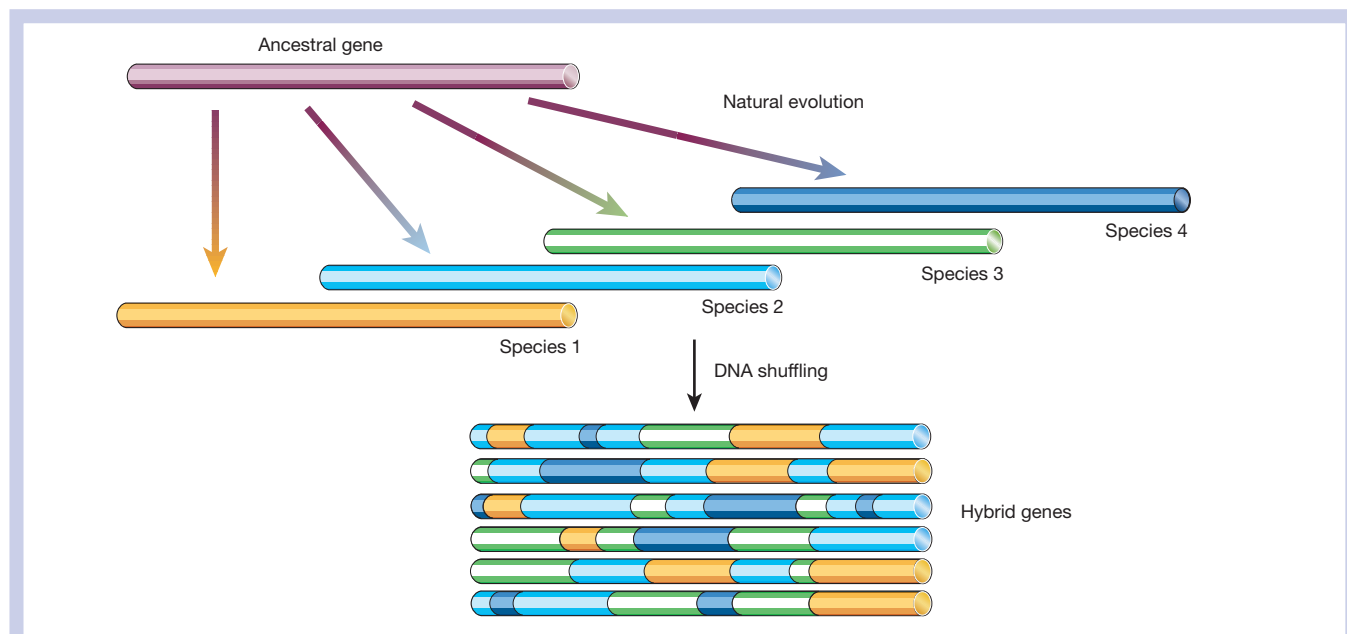
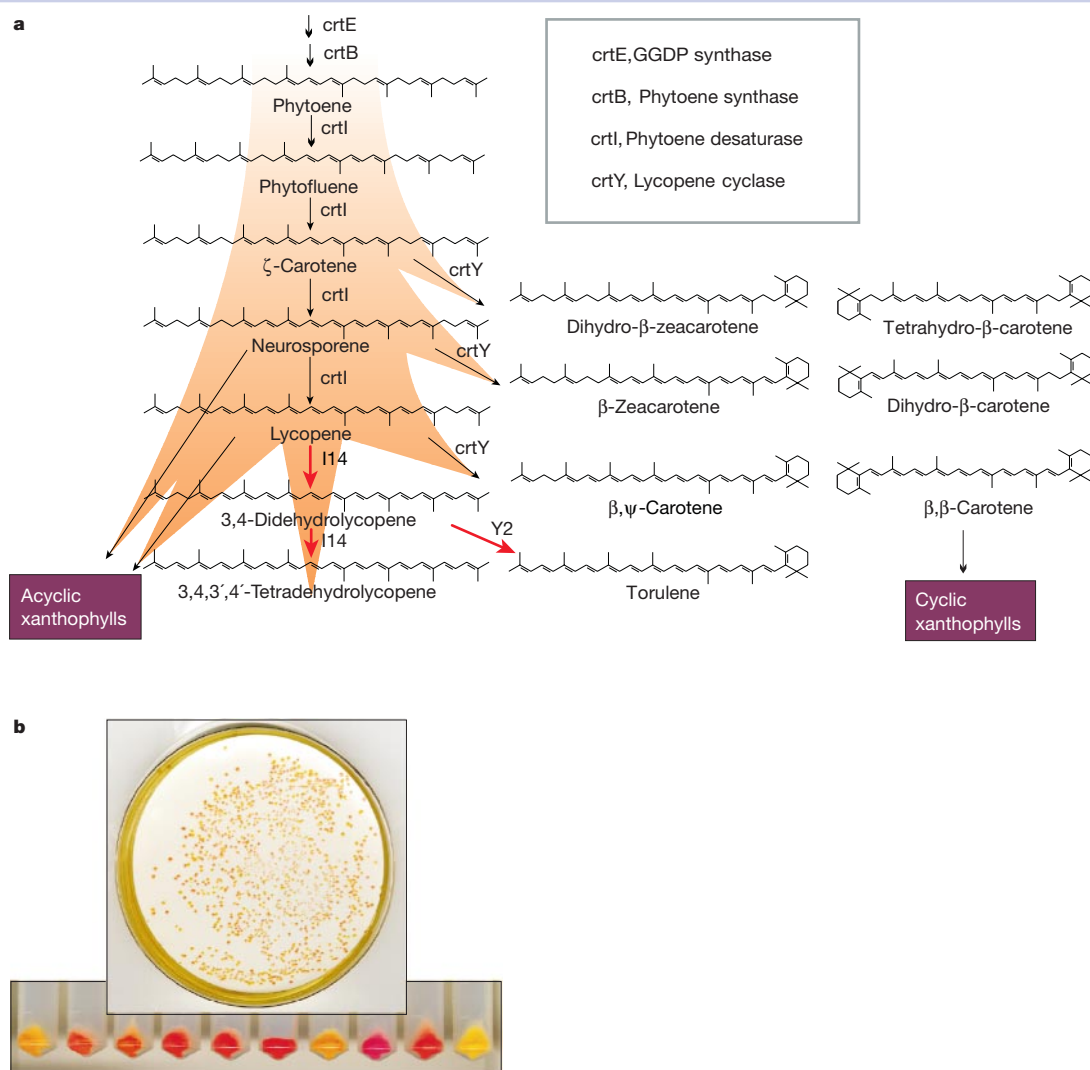


Figure 2 Molecular breeding by DNA shuffling. Diverse gene libraries for laboratory evolution can be created by recombination of related genes³¹. This approach generates highly diverse sequences, but conserves function. Improved or altered enzymes have been identified by screening such hybrid protein libraries.

Figure 3 Evolution of pathways that synthesize carotenoid pigments.

a, Diversity of carotenoid structures produced by molecular breeding of carotenoid biosynthetic genes from *Erwinia* sp.³⁴. C₄₀ carotenoid biosynthesis branches into a variety of pathways to acyclic and cyclic carotenoids for which biosynthetic genes from bacteria have been cloned. Red arrows indicate how molecular breeding of the desaturase extended the central desaturation pathway to generate fully conjugated 3,4,3',4'-tetrahydrolycopene. Subsequent branching of this pathway by a member of a library of shuffled cyclase genes allowed synthesis of torulene, a carotenoid not made by the parent genes and not known in any bacteria. **b**, Coloured bacteria containing shuffled genes synthesize carotenoids not made by the *Erwinia* parents.



E. coli, together with a large library of gene hybrids created by shuffling two versions of a third *Erwinia* gene encoding a desaturase, which normally introduces double bonds into phytoene to make lycopene. Among the thousands of coloured bacterial progeny, they found some that were more yellow and pink than the orange *E. coli* containing the three naturally occurring carotenoid biosynthetic genes. Different members of the bacterial library made one or more of the carotenoids that contained double bonds at the various positions, and all the possible desaturation products were represented.

Combination of the genes that made the pink carotenoid (tetrahydrolycopene) with a new library of mated gene hybrids of a fourth (cyclase) gene generated an even greater variety of coloured bacteria: yellow, orange, pink and bright red (Fig. 3). The bright red cells produce torulene, a carotenoid not made by *Erwinia*, and not known in any bacteria but found in some red yeasts. Yet the pathway created by molecular breeding is not the same as that used by yeast to make torulene.

The combination of gene assembly (pathway engineering) and molecular evolution can solve very complex problems of biological design. By generating efficient pathways to make natural and non-natural products, it can greatly extend the applications of biocatalysis into the discovery and production of new biologically active compounds.

De novo catalyst design

There are likely to be many problems for which natural molecules cannot even offer a suitable starting point for evolution. In some

cases, the whole enzyme frameworks are not suitable because, with many hundreds of amino acids, they are too unwieldy to produce or use in large quantities or, in the case of protein-based drugs, cannot be delivered efficiently to their targets. In other cases, the biological pathways are too cumbersome for practical use. For example, biological oxidation reactions are usually catalysed by large multiprotein complexes and use expensive cofactors that few would consider for an industrial process. In general, the many, sometimes conflicting demands and the contingent nature of evolution means that enzyme structures are not necessarily optimized as chemical reagents for a specific transformation, and there may be much better functional solutions that use completely different sequences and structures. How can one find them?

Some possible routes are evolutionary. Catalytic function can be coaxed out of protein frameworks evolved for different, non-catalytic roles. In one of the first evolutionary approaches to making new biocatalysts, catalytic antibodies or 'abzymes' were generated in response to molecules that mimic the transition state of a reaction³⁵. But the development of commercially useful antibody catalysts has been hampered by their low expression, limited stability and generally low turnover rates, although there are a few notable exceptions³⁶. The basic idea of targeting a transition-state analogue can be extended to generate catalytic activity from other, perhaps more tractable frameworks. But the activities of these new enzymes may still be low, reflecting the fact that transition-state binding is only one aspect of the catalytic process.

Other approaches to designing new protein catalysts use different breeding practices. There are many ways to create molecular diversity beyond point mutation and homologous recombination. Several groups are, for example, investigating nonhomologous recombination of distantly related^{37,38} and even unrelated sequences³⁹ as a means to generate new functional proteins. Others are developing techniques to generate^{40,41} and screen⁴² larger libraries so as to be able to identify rarer but possibly more useful solutions.

But the number of possible protein sequences inevitably dwarfs any existing or even conceivable technology for searching it experimentally. So one must make intelligent choices about what and how to search. This may be where 'rational' design will be crucial: identifying the most likely places to search combinatorially for desired functions.

That rudimentary structure-based designs can be improved through evolutionary tuning is well accepted, if not yet widely practised. This blend of approaches was demonstrated by Altamirano and co-workers, who converted an α/β -barrel enzyme with one activity (indole-3-glycerol phosphate synthase) into another with equally efficient activity (phosphoribosylanthranilate isomerase)⁴³.

Conversely, structure-based computational methods can be used to identify likely sites for evolutionary improvement, thereby supporting the generation of specific 'targeted' libraries and greatly reducing the experimental search. Voigt *et al.* (ref. 44 and unpublished data) have used powerful computational methods¹⁷ to search vast regions of sequence space to identify the most probable solutions to protein-design problems. They use the computational methods where they work best — solving the generic problems of identifying protein sites that are tolerant to mutation or that will tolerate crossover without significant disruption — and the evolutionary methods to find specific solutions within the generic ones.

The ideal would be to specify a catalyst *de novo*: purely from its primary sequence. In principle this should be possible, and indeed the first *de novo* proto-enzymes are now being reported⁴⁵ — although they are not particularly impressive catalysts. Primitive iron- and oxygen-binding sites introduced into the small protein thioredoxin by computational design show varying selectivities in oxidation processes⁴⁵. Such designed sites might be adequate starting points for evolutionary methods.

Conclusions

Biocatalysts need to become predictable and routine tools. At present they are neither, and biocatalyst design is still more of an art than a science. But things are changing. Laboratory evolution methods are now sufficiently robust that improved biocatalysts can be obtained with confidence on a reasonable timescale. Further developments, especially miniaturization and automation of high-throughput screening, will accelerate the acceptance and widespread application of biocatalysis.

Today, evolutionary methods seem the most fertile approach for developing new commercial biocatalysts. But the capabilities of rational design, particularly computational techniques and *de novo* design are expanding too. And emerging design methods that marry the best of the computational and the combinatorial approaches promise to make biocatalysis a key tool for synthetic chemistry in the century ahead. □

- Adam, M. W. W. & Kelly, R. M. Finding and using hyperthermophilic enzymes. *Trends Biotechnol.* **16**, 329–332 (1998).
- Rondon, M. R., Goodman, R. M. & Handelsman, J. The Earth's bounty: assessing and accessing soil microbial diversity. *Trends Biotechnol.* **17**, 403–409 (1999).
- Short, J. Recombinant approaches for accessing biodiversity. *Nature Biotechnol.* **15**, 1322–1323 (1997).
- Bull, A. T., Ward, A. C. & Goodfellow, M. Search and discovery strategies for biotechnology: the paradigm shift. *Microbiol. Mol. Biol. Rev.* **64**, 573–606 (2000).
- Black, M. E., Newcomb, T. G., Wilson, H.-M. P. & Loeb, L. A. Creation of drug-specific Herpes Simplex Virus type 1 thymidine kinase mutants for gene therapy. *Proc. Natl Acad. Sci. USA* **93**, 3525–3529 (1996).
- Encell, L. P., Coates, M. M. & Loeb, L. A. Engineering human DNA alkyltransferases for gene therapy using random sequence mutagenesis. *Cancer Res.* **58**, 1013–1020 (1998).

- Matsumura, I., Wallingford, J. B., Surana, N. K., Vize, P. D. & Ellington, A. D. Directed evolution of the surface chemistry of the reporter enzyme β -glucuronidase. *Nature Biotechnol.* **17**, 696–701 (1999).
- Lanio, T., Jeltsch, A. & Pingoud, A. Towards the design of rare cutting restriction endonucleases: using directed evolution to generate variants of *EcoRV* differing in their substrate specificity by two orders of magnitude. *J. Mol. Biol.* **283**, 59–69 (1998).
- Liu, D. R., Magliery, T. J., Pastrnak, M. & Schultz, P. G. Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins *in vivo*. *Proc. Natl Acad. Sci. USA* **94**, 10092–10097 (1997).
- Babbitt, P. New functions from old scaffolds: how nature re-engineers enzymes for new functions. *Adv. Protein Chem.* **55**, 1–28 (2000).
- Xiang, H., Luo, L., Taylor, K. L. & Dunaway-Mariano, D. Interchange of catalytic activity within the 2-enoyl-coenzyme A hydratase/isomerase superfamily based on a common active site template. *Biochemistry* **38**, 7638–7652 (1999).
- Broun, P., Shanklin, J., Whittle, E. & Somerville, C. Catalytic plasticity of fatty acid modification enzymes underlying chemical diversity of plant lipids. *Science* **282**, 1315–1317 (1998).
- Shanklin, J. Exploring the possibilities presented by protein engineering. *Curr. Opin. Plant Biol.* **3**, 243–248 (2000).
- Shanklin, J. & Cahoon, E. B. Desaturation and related modifications of fatty acids. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 611–641 (1998).
- Wintrop, P. & Arnold, F. H. Temperature adaptation of enzymes: lessons from laboratory evolution. *Adv. Protein Chem.* **55**, 161–226 (2000).
- Jaenicke, R. Stability and stabilization of globular proteins in solution. *J. Biotechnol.* **79**, 193–203 (2000).
- Dahiyat, B. I. & Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **278**, 82–87 (1997).
- Malakoukas, S. M. & Mayo, S. L. Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct. Biol.* **5**, 470–475 (1998).
- Cedrone, F., Méné, A. & Quémener, E. Tailoring new enzyme functions by rational redesign. *Curr. Opin. Struct. Biol.* **10**, 405–410 (2000).
- Arnold, F. H. (ed.) *Evolutionary Protein Design* (Adv. Protein Chem. **55**) (Academic, San Diego, 2000).
- Affholter, J. & Arnold, F. H. Engineering a revolution. *Chem. Britain* **35**, 48–51 (1999).
- Ness, J. E., del Cardayre, S. B., Minshall, J. & Stemmer, W. P. C. Molecular breeding—the natural approach to enzyme design. *Adv. Protein Chem.* **55**, 261–292 (2000).
- Reetz, M. T. & Jaeger, K.-E. Superior biocatalysts by directed evolution. *Top. Curr. Chem.* **200**, 31–57 (1999).
- Liebeton, K. E. Directed evolution of an enantioselective lipase. *Chem. Biol.* **7**, 709–718 (2000).
- May, O., Nguyen, P. & Arnold, F. H. Inverting enantioselectivity by directed evolution of hydantoinease for improved production of L-methionine. *Nature Biotechnol.* **18**, 317–320 (2000).
- Arnold, F. H. Directed enzyme evolution <<http://www.che.caltech.edu:80/groups/fha/Enzyme/directed.html>>
- Oue, S., Okamoto, A., Yano, T. & Kagamiyama, H. Redesigning the substrate specificity of an enzyme by cumulative effects of the mutations of non-active site residues. *Biol. Chem.* **274**, 2344–2349 (1999).
- Yano, T., Oue, S. & Kagamiyama, H. Directed evolution of an aspartate aminotransferase with new substrate specificities. *Proc. Natl Acad. Sci. USA* **95**, 5511–5515 (1998).
- Orencia, M. C. & Stevens, R. C. Structural changes in directed and immune evolved enzymes. *Adv. Protein Chem.* **55**, 227–260 (2000).
- Stemmer, W. P. C. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389–391 (1994).
- Cramer, A., Raillard, S.-A., Bermudez, E. & Stemmer, W. P. C. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288–291 (1998).
- Sandmann, G., Albrecht, M., Schnurr, G., Knörzer, O. & Böger, P. The biotechnological potential and design of novel carotenoids by gene combination in *Escherichia coli*. *Trends Biotechnol.* **17**, 233–237 (2000).
- Cane, D. E., Walsh, C. T. & Khosla, C. Harnessing the biosynthetic code: combinations, permutations and mutations. *Science* **282**, 63–68 (1998).
- Schmidt-Dannert, C., Umeno, D. & Arnold, F. H. Molecular breeding of carotenoid biosynthetic pathways. *Nature Biotechnol.* **18**, 750–753 (2000).
- Patten, P. A. *et al.* The immunological evolution of catalysis. *Science* **271**, 1086–1091 (1996).
- Barbas, C. F. III, List, B., Rader, C., Segal, D. J. & Turner, J. M. From catalytic asymmetric synthesis to the transcriptional regulation of genes: *in vivo* and *in vitro* evolution of proteins. *Adv. Protein Chem.* **55**, 317–366 (2000).
- Ostermeier, M. & Benkovic, S. J. Evolution of protein function by domain swapping. *Adv. Protein Chem.* **55**, 29–78 (2000).
- Siebert, V., Martinez, C. A. & Arnold, F. H. Libraries of hybrid proteins from distantly-related sequences. *Nature Biotechnol.* (in the press).
- Riechmann, L. & Winter, G. Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc. Natl Acad. Sci. USA* **97**, 10068–10073 (2000).
- Plückthun, A., Schaffitzel, C., Hanes, J. & Jermutus, L. In vitro selection and evolution of proteins. *Adv. Protein Chem.* **55**, 367–404 (2000).
- Roberts, R. W. Totally *in vitro* protein selection using mRNA-protein fusions and ribosome display. *Curr. Opin. Chem. Biol.* **3**, 268–273 (1999).
- Olsen, M., Iverson, B. & Georgiou, G. High-throughput screening of enzyme libraries. *Curr. Opin. Biotechnol.* **11**, 331–337 (2000).
- Altamirano, M. M., Blackburn, J. M., Aguayo, C. & Fersht, A. R. Directed evolution of new catalytic activity using the α/β -barrel scaffold. *Nature* **403**, 617–622 (2000).
- Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z.-G. Computational method to reduce the search for directed protein evolution. *Proc. Natl Acad. Sci. USA* (in the press).
- Benson, D. E., Wisz, M. S. & Hellinga, H. W. Rational design of nascent metalloenzymes. *Proc. Natl Acad. Sci. USA* **97**, 6292–6297 (2000).

Acknowledgements

I thank the many talented students and postdocs who have contributed to the development of new biocatalyst engineering tools in my laboratory, and the following organizations for their financial support: the US Office of Naval Research, the US National Science Foundation, the Army Research Office, Maxygen, Inc., The Biotechnology Research & Development Corporation, British Petroleum, Degussa AG and Procter & Gamble Co. I also thank C. Voigt and J. Shanklin for thoughtful comments, and J. Shanklin for Fig. 1.

Industrial biocatalysis today and tomorrow

A. Schmid*, J. S. Dordick†, B. Hauer‡, A. Kiener§, M. Wubbolts|| & B. Witholt*

*Institute of Biotechnology, ETH-Hönggerberg, CH 8093 Zürich, Switzerland (e-mail: bw@biotech.biol.ethz.ch)

†Department of Chemical Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180, USA (e-mail: dordick@rpi.edu)

‡BASF AG, ZHF-B 9, D 67056 Ludwigshafen, Germany (e-mail: bernhard.hauer@basf-ag.de)

§Lonza Biotechnology R&D, CH-3930 Visp, Switzerland (e-mail: andreas.kiener@lonzagroup.com)

||DSM Biotech GmbH, Karl-Heinz-Beckurtsstrasse 13, D 52428 Jülich, Germany (e-mail: marcel.wubbolts@dsm-group.com)

The use of biocatalysis for industrial synthetic chemistry is on the verge of significant growth. Biocatalytic processes can now be carried out in organic solvents as well as aqueous environments, so that apolar organic compounds as well as water-soluble compounds can be modified selectively and efficiently with enzymes and biocatalytically active cells. As the use of biocatalysis for industrial chemical synthesis becomes easier, several chemical companies have begun to increase significantly the number and sophistication of the biocatalytic processes used in their synthesis operations.

Biochemists and microbiologists have long seen biocatalysis as an area with great promise for chemical synthesis, but industrial applications have been modest. In this review article we sketch the current state of industrial biocatalysis in several European industries and look ahead to new processes that are likely to develop, based on current academic and industrial research. Two parallel developments are apparent in industry: chemical industries are hiring increasing numbers of life scientists, and organic chemists are beginning to embrace biocatalysis as a tool in new and difficult syntheses. This will lead to more industrial applications of biocatalysts.

Work during the past decade has shown that there are surprisingly few barriers to the use of enzymes and whole cells as biocatalysts in organic synthesis^{1,2}. Isolated enzymes are typically used for hydrolytic or isomerization reactions. Whole cells are often used for synthetic reactions that require cofactors which must be regenerated, because although cofactor regeneration *in vitro* is possible, it is generally easier and less expensive to regenerate cofactors in metabolically active cells. Both isolated enzymes and whole cells are used in industry today, and are an active area of research.

Enzymes are remarkable catalysts: capable of accepting a wide array of complex molecules as substrates, and exquisitely selective, catalysing reactions with unparalleled chiral (enantio-) and positional (regio-) selectivities. As a result, biocatalysts can be used in both simple and complex transformations without the need for the tedious blocking and deblocking steps that are common in enantio- and regioselective organic synthesis. Such high selectivity also affords efficient reactions with few by-products, thereby making enzymes an environmentally friendly alternative to conventional chemical catalysts.

These attributes have resulted in myriad applications, especially in the food and pharmaceutical industries where high reaction selectivity on complex substrates is critical. Examples include the production of high-fructose corn syrup, by the action of xylose isomerase³ which catalyses the isomerization of D-glucose to D-fructose, and the preparation of semisynthetic penicillins catalysed by penicillin

amidase⁴. Selective catalysis is now also becoming a requirement for the chemical industry, and recent advances in enzymatic catalysis have been extended to the synthesis of speciality chemicals and polymers^{5,6} and of some bulk chemicals. For example, peroxidases are used industrially to catalyse the synthesis of phenolic resins for use as replacements of conventional phenol-formaldehydes⁷, and nitrile hydratase is used to catalyse the hydration of acrylonitrile into acrylamide⁸. In both cases, nearly quantitative conversion of the reactants into products is obtained and under reaction conditions far milder than their chemical counterparts. Most commercial enzymatic processes today share several attributes, including high product concentrations and productivities, no undesirable by-products, and enzymes that do not require expensive cofactors.

Future biocatalytic processes generally will not be limited by the available technology or the nature of the substrates and the products. Instead, the feasibility of new biocatalytic processes will often be determined by the availability of the biocatalyst, the search for which is described in the accompanying papers in this issue by Walsh (pages 226–231), Arnold (pages 253–257) and Khosla and Harbury (pages 247–252). Consequently, a growing number of companies sees biocatalysis as an interesting option. As individual industries develop relevant experience, industrial biocatalysis will grow rapidly.

The biocatalysis cycle

Biocatalytic processes differ from conventional chemical processes, owing mainly to enzyme kinetics, protein stability under technical conditions and catalyst features that derive from their role in the cell's physiology, such as growth, induction of enzyme activity or the use of metabolic pathways for multistep reactions. In the laboratory, new biocatalytic reactions often originate with new enzyme activities. For applications, a more rational approach is needed. The starting point will usually be a product, which can perhaps be produced by one of several possible biocatalytic reactions that convert suitable substrates to the desired product. Figure 1 illustrates the development of such biocatalytic processes. One or more biocatalysts must

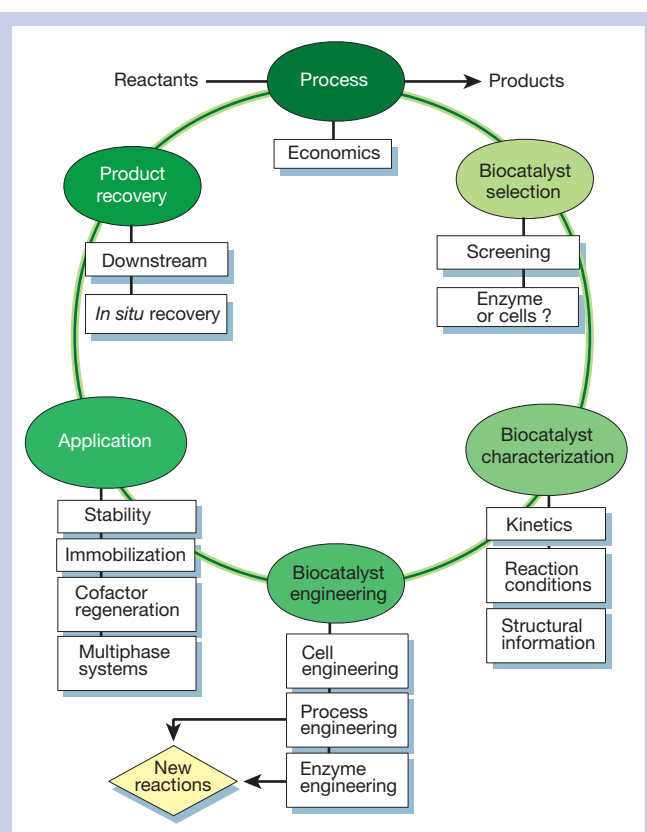


Figure 1 The biocatalysis cycle.

be identified or developed, a process must be set up, and the resulting bioconversion will ultimately have to be economically feasible. The development of such a process requires the input of many different specialists. Limiting aspects of the biocatalytic process are improved in an iterative manner, gradually leading to an efficient industrial process. In setting priorities for improvements at each process step, a detailed understanding of the costs and improvement potential of each of the partial steps in a process is vital.

The economic feasibility of a biocatalytic process depends on several factors (Fig. 1). Depending on the type of biocatalyst to be used, specific reactor and hardware configurations are needed (summarized in ref. 9). In addition, biocatalytic processes are typically highly heterogeneous. In theory, this would necessitate specific designs of the catalyst–hardware interface¹⁰. But in practice, a limited number of hardware designs is found today in large industrial processes, allowing the application of biocatalysts based on only a few concepts. In analogy to chemical processes, most biocatalysts are used in immobilized form as heterogeneous catalysts that can be recovered and reused. There are also processes, however, based on homogeneously suspended cells or enzymes, that are sufficiently inexpensive to permit single use, without recovery or reuse. In fact, several speciality chemical companies now use living cells as catalysts for reactions such as specific coenzyme-dependent oxidoreductions, as described below.

The biocatalyst

New processes can be based on the availability of an interesting new enzyme, or on the identification of desired products, after which a biocatalyst is then selected that permits conversion of available reactants. Such an enzyme might be available commercially, or it might have been described in the literature. Alternatively, it will be necessary to screen for organisms or enzymes that carry out the desired reaction, or completely new enzyme activities will be developed by protein design or directed evolution (see review in this issue by Arnold, pages 253–257).

For conversions that do not require regeneration of coenzymes, such as isomerization or hydrolysis reactions, both enzymes and whole cells can be selected. But when cofactors are required, whole cells are favoured because they enable cofactor regeneration. Reaction conditions for optimal enzyme function, high reactivities and long catalyst lifetimes are selected based on biocatalyst characteristics. The biocatalyst and the biocatalytic process are engineered for best performance — at the level of the enzyme (protein engineering for better activity, improved substrate range, enzyme stability), the host cell (solvent resistance, substrate import and product export, elimination of side-reactions), or the process. The biocatalyst (enzymes or cells) may be immobilized, and cofactors regenerated for coenzyme-dependent enzymes. The reaction medium, which may consist of an aqueous phase, an organic phase or a two-liquid-phase system, will be optimized to dissolve substrates and products while maintaining enzymatic activity.

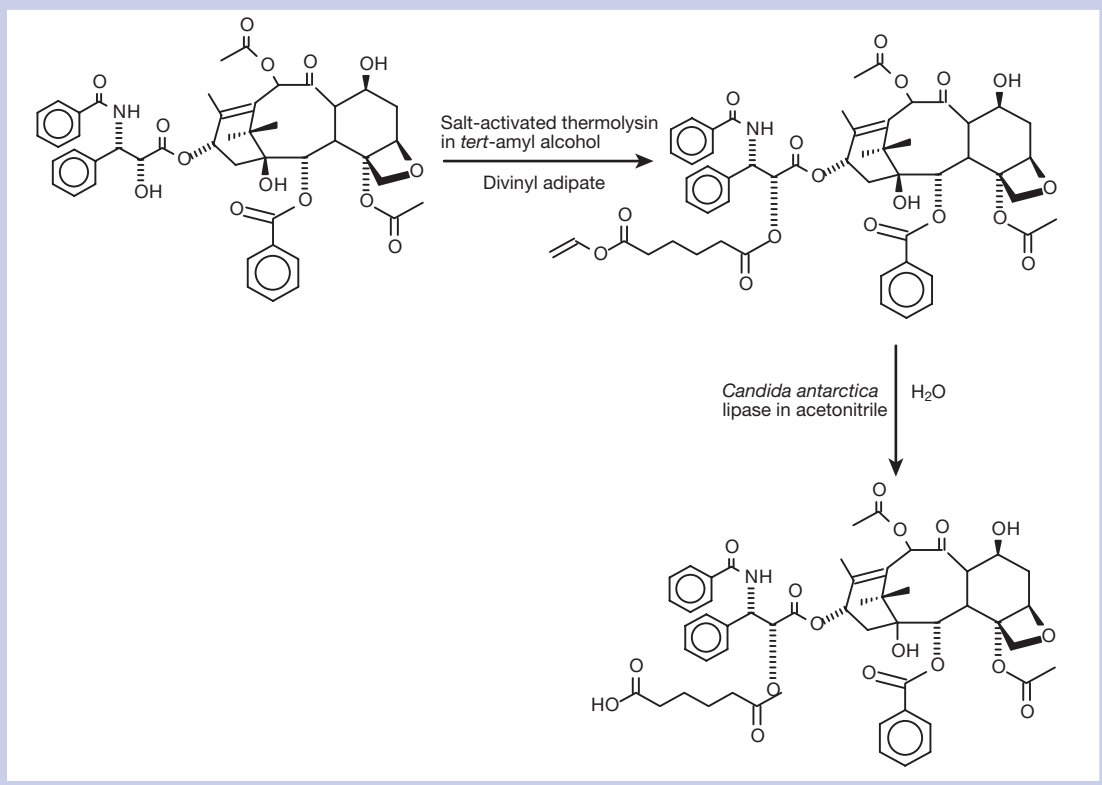
Enzyme catalysis in organic solvents

The rapid growth of biocatalysis is a direct result of research and development in two key technologies: protein engineering, including molecular evolution^{11,12}, and enzyme engineering. Whereas the former provides enzymes with altered structure, function and selectivity, particularly in aqueous media, the latter, especially involving engineering of the enzyme microenvironment, provides striking improvements in nonaqueous environments. Indeed, it is now well known that enzymes do function in organic solvents, and many in neat (pure) solvents or in supercritical fluids in the absence of added water (see refs 13–15 and review in this issue by Klivanov, pages 241–246). Such an environment yields many potential advantages, including higher substrate solubility, reversal of hydrolytic reactions and modified enzyme specificity, which result in new enzyme activities that previously were only possible using genetic modifications or complex multistep pathways within whole cells. As a result, applications of enzymatic catalysis in organic solvents range from chiral resolution of pharmaceuticals, chemicals and their intermediates¹⁶ to enantio- and regioselective polymerization¹⁷.

Despite the advantages of nonaqueous conditions for biocatalytic transformations, enzymes nearly universally display low catalytic activities in these environments compared with native aqueous solutions. Nonetheless, recent developments have shown that biocatalysts can be engineered to function in neat organic solvents with activities and selectivities that are consistent with their aqueous-based counterparts. For example, subtilisin Carlsberg suspensions (subtilisin and other enzymes are insoluble in nearly all organic solvents) prepared simply by lyophilizing an aqueous preparation (see review by Klivanov, pages 241–246) in the presence of non-buffer salts¹⁸ yield rate enhancements of more than 20,000-fold¹⁹. The mechanism of this activation has not been elucidated, although it is strongly dependent on the specific salt used. Specifically, kosmotropic (order-promoting) salts are expected to stabilize the folded form of enzymes during the lyophilization process as well as in the nonaqueous reaction medium, thereby leading to higher enzymatic activities in such solvents²⁰. In addition to subtilisin, a number of other enzymes are activated by this technique, including those with very different catalytic mechanisms. Activation has also been achieved by the addition of crown ethers²¹, transition-state analogues²², and substrates and substrate mimics²³.

Activated biocatalyst preparations have found direct application in the pharmaceutical industry, where salt-activated biocatalysts have been used to synthesize a library of paclitaxel (taxol) derivatives²⁴. The bacterial protease thermolysin was found to acylate selectively the 2'-hydroxyl of taxol in *t*-amyl alcohol (Fig. 2). Yields of the 2'-acyl derivatives approached 100% using KCl-activated thermolysin. For the specific acylation with divinyladipate, a taxol 2'-vinyladipate was generated, which served as the acyl donor for *Candida antarctica* lipase-catalysed hydrolysis of the terminal vinyl ester. The resulting taxol 2'-adipic acid derivative was nearly 1,700 times more soluble in water than the native taxol, a result of critical

Figure 2 Synthesis of a library of paclitaxel (taxol) derivatives.



importance in the design of taxol prodrugs with increased bioavailability.

In the presence of low concentrations of a suitable surfactant, enzymes are able to dissolve in hydrophobic organic solvents, where they remain remarkably active and with secondary and tertiary structures nearly identical to that measured in water²⁵. For example, in peptide synthesis, subtilisin Carlsberg and α -chymotrypsin were over 1,000-fold more reactive than their native suspended counterparts in suitable organic solvents. Solubilized enzymes have been used to generate 'biocatalytic plastics' (Fig. 3), wherein enzymes are incorporated into growing vinyl polymers to yield homogeneous immobilized preparations^{26,27}. This technique enables the biocatalyst to be used in a form suitable for a specific function. Thus, suspension polymerization could be used to yield biocatalytic plastic beads of controlled sizes, for example as catalysts in packed-bed reactors. Thin-film formation could be pursued to give biocatalytic paints, coatings and films for applications ranging from antifouling coatings (for example to prevent surface protein and cell adhesion) to affinity materials for use in the synthetic, diagnostic and medical arenas.

The development of 'solvent-free' systems has become of interest recently as a more environmentally benign technique for catalysing reactions that cannot be performed in aqueous solutions. For example, enzymatic polyester synthesis has been performed in neat solutions of diols and diesters²⁸. Rapid polymer growth is obtained with nearly quantitative conversions.

Whole-cell catalysis with toxic solvents

A number of potentially interesting biocatalytic conversions being investigated today involve apolar substrates and products, such as aliphatic, aromatic and heterocyclic compounds. Such compounds are generally insoluble in water, and often they are toxic to whole cells^{29–31}. Thus they cannot simply be added to an aqueous medium for whole-cell transformation. Several technical solutions to this problem have been developed.

One promising approach is to use two-liquid-phase media: an aqueous phase that contains the growing cells, and an apolar solvent that contains the substrate and newly formed product^{32–34}. Biocatalysis in emulsions is a well established technology in research

laboratories^{35–37} and is expected to be equally applicable on larger scales^{38–40}. Moreover, solvent-based processes enable the use of well established, industrial downstream-processing techniques. For two decades from the 1970s these systems were used with various catalytically active natural hosts, often pseudomonads, but lately recombinants harbouring genes for appropriate enzymes are gaining ground. Although not particularly solvent-resistant, *Escherichia coli* has been used effectively as a biocatalysis host. Examples include the oxidation of alkanes^{34,35}, aromatics such as toluene and styrene derivatives⁴¹, and heterocycles⁴².

A related engineering approach for processes based on multi-phase reaction media is gas-phase biocatalysis⁴³. Enzymes or intact cells form a solid phase and reactants are dissolved in the gas phase. This concept is suitable for reactants that can be brought into the gas phase at operating conditions, and has the advantage of simple downstream processing and efficient mass transfer, while the use of toxic or inhibiting solvents can be avoided.

A third approach to handling toxic compounds is to feed these into the bioreactor at limiting rates, so that such substrates are transformed without first accumulating, thus maintaining very low and non-toxic substrate concentrations. Products can be removed essentially instantaneously by continuous extraction techniques, either in suitable apolar solvents or by adsorption onto solid beds⁴⁴.

These techniques have now matured to a point where they can be used routinely to carry out whole-cell biotransformations on a wide range of substrate–product combinations. Perhaps surprisingly, *E. coli* may well be the preferred host. Such transformations are limited only by the availability of the required enzyme systems. If these can be found (from available strains, screening or selection) or generated (by protein engineering or directed evolution), a suitable whole-cell biocatalyst can be constructed^{41,45}. By using highly water-insoluble solvents such as hexadecane, two-liquid-phase systems can be developed that contain 10–50% (by volume) apolar phase in an aqueous medium, and cause no damage to the *E. coli* host^{39,46,47}. Recombinant strains can be grown to densities of 10–40 g l⁻¹ (dry mass), either in batch, fed-batch or continuous-cultivation systems³⁰. Overall enzyme activities for monooxygenases typically vary from 100 to 500

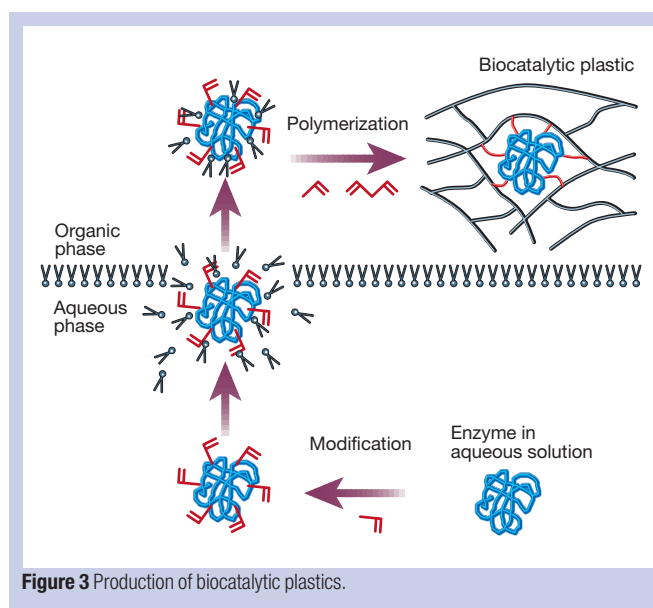


Figure 3 Production of biocatalytic plastics.

international units (micromoles converted per minute) per litre of aqueous medium, and these activities can be maintained for 10–20 hours^{48,49}. These numbers lead to space-time yields of 5–25 tons of product per cubic metre aqueous medium per year, for typical small molecules with relative molecular masses of 100–200. Higher numbers might be reached, based on enzyme activities of 30–100 international units per gram cell (dry mass) and cell densities of 10–20 g l⁻¹ (ref. 48). Other parameters would, however, then become limiting, such as the practical volumetric energy and oxygen input and the heat removal capacity that can be achieved in large-scale industrial reactors.

Product extraction and purification

Most biocatalytic processes still operate in aqueous environments and often result in low product concentrations. This necessitates the development of special methods for product extraction. In contrast to industrial bulk chemical syntheses, bioprocesses are usually operated in batch mode; continuous processes are still the exception. Recovery of product from the reaction medium generally follows the biotransformation step as a separate batch operation, after separation of biomass from the reaction mixture either by centrifugation or filtration.

Even techniques as simple as precipitation of insoluble reaction products are used, for instance in the process for acrylamide production from acrylonitrile at product concentrations as high as 600 g l⁻¹ (ref. 50). Such *in situ* product-recovery processes, in which the product is separated from the reaction mixture during the biotransformation, have distinct advantages and are becoming popular (reviewed in ref. 51). These *in situ* product-recovery techniques are based on physical and chemical parameters of the reactants and range from crystallization through distillation to solid-phase or liquid-phase extraction. In general, *in situ* product recovery overcomes traditional limitations of biocatalytic processes such as substrate or product inhibition and product decay in aqueous reaction mixtures.

In the case of biocatalysis in mixed aqueous and apolar solvent phases, which is relevant to bioconversions of hydrocarbons and other apolar compounds⁵², the phases are separated after the biocatalytic step, followed by subsequent product adsorption, liquid extraction or distillation, generally from the apolar phase^{40,53,54}. These steps may in turn affect the selection of the reaction medium and reactor configuration, so that a biocatalytic process may be developed in several iterations through the biocatalysis cycle of Fig. 1. A number of recently developed processes are listed in Table 1.

Waste processing

The major waste products of aqueous bioprocesses are waste water, salts and biomass. Organic solvents represent minor waste products if used in emulsion processes or reactions in pure organic solvents:

they can be recycled, apart from a minor fraction obtained as insoluble emulsion. The main waste products are treated in conventional industrial waste-water treatment systems. For processes based on recombinant, viable microorganisms, the treatment of biomass includes inactivation of the genetically engineered cells.

Biocatalytic processes at BASF

BASF (Germany) has several biotechnological processes in place for the large-scale manufacture of a number of important optically active building blocks and their derivatives. Critical for the development of these processes was a thorough understanding of the catalytic mechanism of the enzymes involved, such as lipases.

Lipase-based processes

Lipids are compounds that can be extracted in chloroform–methanol mixtures. They consist primarily of fatty-acid chains, which are linked by ester bonds to an alcohol or polyol backbone. Lipases belong to a large class of enzymes that hydrolyse the ester bond between the fatty-acyl side chains and the lipid backbone. Understanding the catalytic cycle of lipases has been of significant importance to their widespread use in different biotechnological applications. The lipase active site is composed of three different residues: serine, histidine and aspartate or glutamate.

The hydrolysis of an ester involves an acyl enzyme complex. The catalytic cycle starts by nucleophilic attack of the hydroxyl group of the serine side chain on the carbonyl carbon atom of the ester bond. The complex is resolved by the nucleophilic attack of water, the fatty acid is liberated and the enzyme is regenerated (Fig. 4).

As lipases are active in organic solvents, water can be replaced by other nucleophiles such as alcohols. The result of this reaction is a transesterification. For racemic alcohols only one enantiomer may be acylated, thereby leading to enantioselective transformations (see accompanying reviews in this issue by Walsh, pages 226–231, and Koeller and Wong, pages 232–240). Suitable acyl donors are vinyl esters, anhydrides or diketene. The reaction is irreversible and the separation of the remaining alcohol and the newly formed ester is simple. This principle is now used in many reactions to produce enantiomerically pure alcohols. BASF has recently extended its chiral synthesis capabilities to the enzymatic resolution of racemic alcohols. Some of the core technologies at BASF, such as condensation or hydrogenation of ketones, give access to the racemic starting compounds. A broad range of enantiomerically pure alcohols can now be offered (Fig. 5a).

Amines might also be used as nucleophiles⁵⁵. Racemic amines are efficiently resolved using ethylmethoxyacetate as acylating agent (Fig. 5b). Using ethylmethoxyacetate the initial reaction rate is more than 100 times faster compared with butylacetate. The reason for this activating effect of the methoxy group is probably an enhanced carbonyl activity induced by electronegative α -substituents. Excellent yields and selectivity and minimal amounts of enzyme

Figure 4 Reaction mechanism of lipase biocatalysis. The nucleophilic attack on carbonyl functionalities serves as target for reaction engineering approaches.

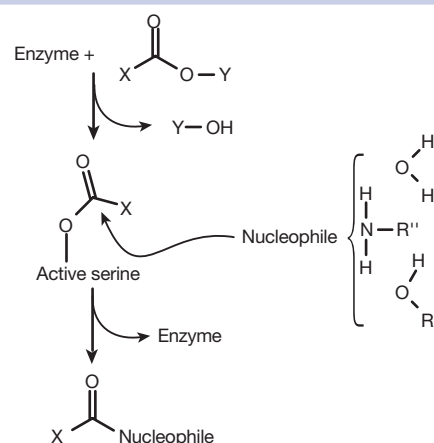


Table 1 Recently developed biocatalytic systems at several chemical companies

Figure	Product	Substrate	Reaction	Biocatalyst	Enzyme	Organism*	Scale (Tons yr ⁻¹)	Yield	Applicability	Source
Amides, alcohols, acids										
5a	Enantiopure alcohols	Racemic alcohols	Resolution	Enzymes	Lipases		Thousand	Excellent	Broad spectrum of alcohols	BASF
5b	<i>R</i> -Amide; <i>S</i> -amine	Racemic amines	Resolution	Enzymes	Lipases		Several hundred	Excellent	Broad spectrum of amines	BASF
5c	<i>R</i> -Mandelic acid	Racemic mandelonitrile	Hydrolysis	Enzymes	Nitrilases		Several	Excellent		BASF
Amino acids, penicillins										
6a	Non-proteinogenic L-amino acids	Racemic amino-acid amides	Kinetic resolution	Enzymes	Amidases	<i>P. putida</i> , <i>M. neoaurum</i> , <i>O. antropii</i> , rec. <i>E. coli</i>	Few to several hundred		Broad spectrum of L- and D-amino acids	DSM
6b	L-Aspartic acid	Fumaric acid	Addition of ammonia	Enzymes	Aspartic acid ammonia lyase	<i>E. coli</i>	Thousand		L-Aspartic acid	DSM
6b	Aspartame (L- α -aspartyl-L-phenylalanine methyl ester)	N-protected L-aspartic acid, D/L-phenylalanine methyl ester	Selective coupling	Enzymes	Thermolysine	<i>B. subtilis</i>	Thousand		Aspartame	DSM
6c	6-Aminopenicillanic acid (6-APA)	Penicillin G/V	Hydrolysis	Enzymes	Penicillin acylase	<i>E. coli</i>	Thousand		6-APA	DSM
6c	Semisynthetic penicillins	6-Aminopenicillanic acid	Selective coupling	Enzymes	Acylases	<i>E. coli</i> , <i>P. putida</i>	Few to several hundred		Semisynthetic penicillins and cephalosporins	DSM
N-Heterocyclic compounds										
7a	6-Hydroxynicotinic acid	Niacin	Addition of water	Whole cells	Niacin hydroxylase	<i>A. xylosoxidans</i> LK1	Few	65 g l ⁻¹	Insecticides	Lonza
7b	5-Hydroxypyrazine-carboxylic acid	2-Cyanopyrazine	Addition of water	Whole cells	Nitrilase/hydroxylase	<i>Agrobacterium</i> DSM 6336	Development product	40 g l ⁻¹	Antituberculosis drugs	Lonza
7c	6-Hydroxy-S-nicotine	(S)-Nicotine	Addition of water	Whole cells	Hydroxylase	<i>A. oxydans</i> NRRL-B-3603	Development product	30 g l ⁻¹		Lonza
7c	4-[6-Hydroxypyridin-3-yl]-4-oxobutrate	(S)-Nicotine	Complex reaction	Whole cells	Several	<i>Pseudomonas</i> sp. DSM 8653	Development product	15 g l ⁻¹	Analogues of Epibatidine	Lonza
Non-proteinogenic amino acids										
7d	S-Piperazine-2-carboxylic acid	(R,S)-Piperazine-2-carboxylic acid	Selective amidase	Whole cells	Stereospecific amidases	<i>K. terrigena</i> DSM 9174	Development product	9 g l ⁻¹	Various bioactive compounds	Lonza
7e	5-Methylpyrazine-2-carboxylic acid	2,5-Dimethylpyrazine	Selective oxidation of a methyl substituent on an aromatic N-heterocycle (pyrazine)	Whole cells	Xylene oxidation pathway	<i>P. putida</i> ATCC 33015	Several	20 g l ⁻¹	Synthon for antilipolytic drug	Lonza

*Organism: *A. oxydans*, *Arthrobacter oxydans*; *A. xylosoxidans*, *Achromobacter xylosoxidans*; *B. subtilis*, *Bacillus subtilis*; *E. coli*, *Escherichia coli*; *K. terrigena*, *Klebsiella terrigena*; *M. neoaurum*, *Mycobacter neoaurum*; *O. antropii*, *Ochrobactrum antropii*; *P. putida*, *Pseudomonas putida*.

characterize this new process, which has been used by BASF since 1993. The products, the *R*-amide and the *S*-amine, can be recovered and separated by distillation and have high chemical and optical purities.

The above process is applicable to a broad spectrum of amines, which are of considerable interest as chiral building blocks or as auxiliaries for the syntheses of bioactive ingredients. Several chiral amines (Fig. 5b) are now being produced in scales up to several hundred tons in a multiproduct facility. BASF is building a second-generation plant using this new technology, which is due to begin production in 2001 and will be operated according to cGMP (current Good Manufacturing Practice) standards. Another optically active amine that will be produced by this process is *S*-methoxyisopropylamine. It is a building block for the optically active corn herbicide Frontier x2. A dedicated plant for this amine will go on-stream in mid-2001 with an annual capacity of 2,000 tons.

For the success of lipases as industrial biocatalysts it has been important that they be available in commercial quantities, highly stable and active in organic solvents, and that they have a broad substrate range. Besides their usefulness in the production of alcohols and amines, lipases can be used for other reactions. For example, enolesters can be stereoselectively protonated and other reactions such as the cleavage of oxime esters, C–C bond formation and synthetic Michael reactions (the nucleophilic addition of a carbanion to an α,β -unsaturated carbonyl compound) may be achievable. But significant further improvements are necessary to facilitate these reactions⁵⁶.

Nitrilase-based processes

An interesting class of building blocks are amino- and hydroxycarboxylic acids. Such compounds can be synthesized stereoselectively in high enantiomeric excess by nitrilases, a class of enzymes that catalyse the hydrolysis of a nitrile to the corresponding ammonium salt of a carboxylic acid (Fig. 5c). Again this class of enzymes fits nicely in existing product lines and synthesis capabilities at BASF. One such sector involves hydrogen cyanide chemistry, which provides access to the starting materials for subsequent enzymatic bioconversion to chiral α -hydroxy- or α -aminocarboxylic acids with nitrilases. For the α -hydroxynitriles (Fig. 5c, X = OH), both enantiomers can be used because cyanohydrins are known to racemize in aqueous solution through an equilibrium between aldehyde and hydrogen cyanide. By choosing the right pH conditions, quantitative enzymatic conversions can be achieved. Nitrilases therefore provide an extension of industrial hydrogen cyanide chemistry, giving access to a wide range of new intermediates. Microorganisms having nitrilase activity can be enriched from nature by using nitriles as the only nitrogen and/or carbon source in the growth medium. To find three new nitrilases, about 5,000 strains from enriched cultures and isolates were tested. One is currently used in a pilot plant to make *R*-mandelic acid on a multiton scale (Fig. 5c)⁵⁷.

Biocatalytic processes at DSM

The Dutch chemical company DSM has a long history in biocatalysis, starting with the pioneering work of Boesten *et al.*⁵⁸ in the mid 1970s

on the enzymatic resolution of D/L-phenylglycinamide into D-phenylglycinamide and L-phenylglycine using hog leucine aminopeptidase and an L- α -aminoacylamidase from *Pseudomonas putida*⁵⁹. Currently, DSM uses biocatalysis, biotransformation and fermentation technologies in addition to chemical methods to produce advanced intermediates for the custom manufacturing arena^{5,60,61}. Some examples of processes that make use of enzymes and that have been commercialized by DSM are listed below.

Production of non-proteinogenic L-amino acids

The demand for non-proteinogenic L-amino acids as building blocks for the synthesis of pharmaceutical compounds is considerable and these amino acids are not accessible by fermentation, in contrast to most of the 20 proteinogenic amino acids that are incorporated into polypeptides. The DSM process for the production of L-amino acids is based on the resolution of racemic amino-acid amides (**1** and *rac*-**1** in Fig. 6a), which are conveniently synthesized from aldehydes using the Strecker synthesis. Resolution is achieved with amino-acid amide hydrolases (amidases). Although the enzymes are strictly L-selective, both D- and L-amino acids (respectively **L-2** and **D-2** in Fig. 6a) can be accessed and the undesired enantiomer can be recycled.

The biocatalytic amidase toolbox of DSM consists of enzymes from *P. putida*, *Mycobacter neoaurum* or *Ochrobactrum anthropii*, which vary in substrate range, specific activity and enantioselectivity⁶⁰. As a result, a wide range of amino-acid amides can be resolved (*rac*-**1** in Fig. 6a)^{5,62–67}. The genes encoding the amidases have been isolated and overproduction of the enzymes in *E. coli* has successfully been carried out, resulting in much improved biocatalysts⁶⁰.

Production of aspartame

The low-calorie sweetener aspartame (L- α -aspartyl-L-phenylalanine methyl ester; **5** in Fig. 6b) is produced on a kiloton scale by Holland Sweetener Company, a joint venture of Tosoh and DSM. The aspartame process uses a proteolytic enzyme, thermolysin, to catalyse the formation of the dipeptide from N-protected L-aspartic acid (Z-Asp; **Z-3** in Fig. 6b) and D/L-phenylalanine methyl ester (*rac*-**4** in Fig. 6b)⁵. L-aspartic acid (**3**) for the production of Z-Asp is accessible by biocatalysis using aspartic acid ammonia lyase (aspartase), which catalyses the addition of ammonia to fumaric acid. Subsequently, the protective group is attached, resulting in Z-Asp⁶⁸.

The coupling of the two amino acids Z-Asp and D/L-phenylalanine methyl ester by thermolysin is not only the reversal of the hydrolytic reaction that is common to the enzyme, but also displays a remarkable selectivity. Out of the four possible conformations, taking both the α - and β -carboxyl group of Z-Asp into account, thermolysin selectively couples L-phenylalanine methyl ester to the α -carboxyl group of Z-Asp, thereby generating Z-aspartame (**Z-5** in Fig. 6b). The

Z-group is removed by catalytic hydrogenation and the remaining D-phenylalanine methyl ester can be racemized and recycled.

Green alternatives to produce semisynthetic penicillins and cephalosporins

DSM produces penicillin G/V by fermentation using *Penicillium chrysogenum* strains, which have been improved by classical strain improvement as well as genetic engineering. Most of the penicillin (for example, penicillin G; **6** in Fig. 6c) is converted to 6-aminopenicillanic acid (**6-APA**; **7** in Fig. 6c), which serves as a backbone for the synthesis of semisynthetic penicillins, providing a range of penicillin variants with differing antibiotic characteristics. The 6-APA core structure can be extended by chemical means or by the action of an enzyme to a core structure suitable for synthesis of semisynthetic cephalosporins, such as 7-aminodesacetoxycephalosporanic acid (**7-ADCA**; **8** in Fig. 6c). The coupling of side chains to the core structures generates derivatives of penicillins and cephalosporins that have traditionally been produced using complex solvent-based chemistry, typically conducted at temperatures as low as $-80\text{ }^{\circ}\text{C}$ to preserve the labile β -lactam ring.

Current processes use penicillin acylases to remove the side chain from penicillin G/V, providing a 'green' route to 6-APA⁶⁹. For example, the β -lactam ring of penicillin G is converted chemically or enzymatically to phenylacetyl-7-ADCA, which is also accepted by the acylase, thus producing the nucleus for the expandase-catalysed synthesis of semisynthetic cephalosporins⁷⁰.

D-phenylglycine and D-*p*-hydroxyphenyl glycine, which are the side chains that are attached to the 6-APA or 7-ADCA back-bone in the semisynthetic antibiotics ampicillin, amoxycillin, cephalexine and cephadroxy (respectively **9**, **10**, **11** and **12** in Fig. 6c), are also produced by DSM. The coupling of the side chain to the nucleus is largely performed chemically, but DSM has recently completed the development of a biocatalytic process using another penicillin acylase^{4,5,71,72}. Under certain conditions of pH and by controlling precursor and product concentrations, the coupling of the two molecules can out-compete the hydrolysis reaction catalysed by the same enzyme.

Biocatalytic processes at Lonza

The Swiss custom-manufacturing company Lonza specializes in the production of many chemical compounds, including *N*-heterocycles. Lonza has developed a series of biocatalytic routes for the production of certain functionalized *N*-heterocycles where chemical synthesis is inefficient. The biotransformations described here can be divided into three types: the selective ring hydroxylation of aromatic *N*-heterocycles, the kinetic resolution of racemic *N*-heterocyclic carboxamides to the corresponding enantiomerically pure

Figure 5 Recently developed biocatalytic systems at BASF. Lipases are successful biocatalysts used in the synthesis of **a**, enantiomerically pure alcohols, and **b**, chiral amines. **c**, Nitrilases are used in the synthesis of amino- and hydroxycarboxylic acids.

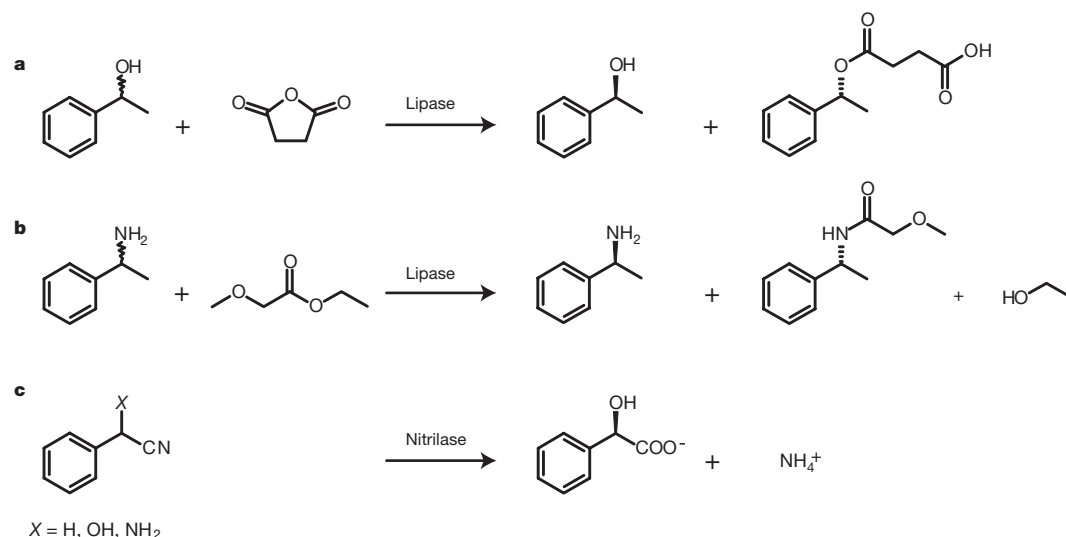
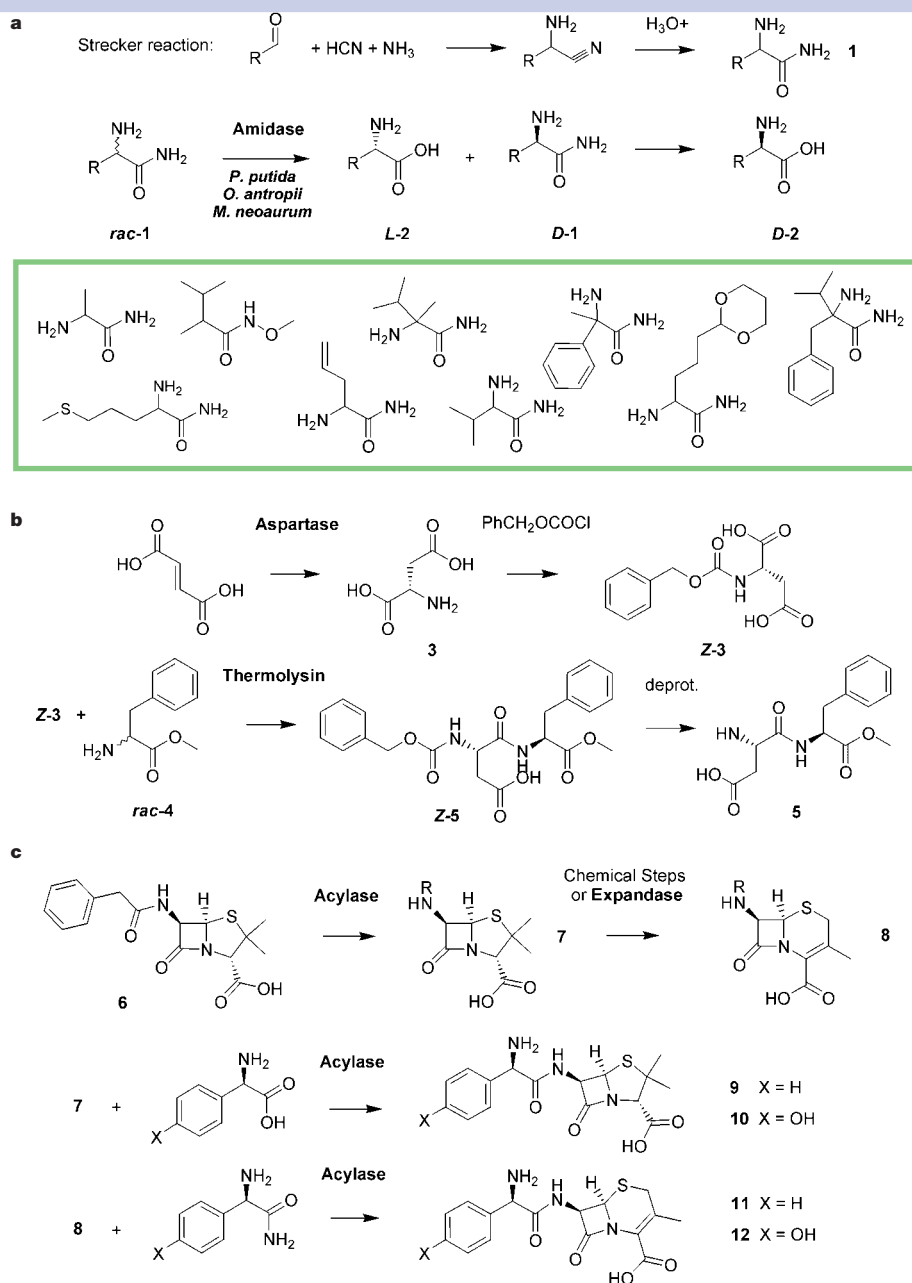


Figure 6 Biocatalytic processes at DSM. **a**, Amidase process for production of non-proteinogenic amino acids. **b**, Biocatalytic steps for the production of aspartame. **c**, Acylases for the production of semisynthetic penicillins and cephalosporins.



carboxylic acids with stereospecific amidases, and the selective oxidation of alkyl groups on aromatic *N*-heterocycles.

6-Hydroxynicotinic acid and 5-hydroxypyrazinecarboxylic acid

6-Hydroxynicotinic acid (**1** in Fig. 7a), a derivative of niacin (**2** in Fig. 7a), serves as a versatile building block predominantly in the synthesis of modern insecticides. The chemical synthesis of 6-substituted niacin results in the formation of by-products which make the cost prohibitively high. Lonza has detected microorganisms growing on niacin capable of accumulating up to 65 g l^{-1} 6-hydroxynicotinic acid at an overall yield of 90% (ref. 73). Remarkably, at niacin concentrations greater than 10 g l^{-1} the second enzyme of the pathway (6-hydroxynicotinate hydroxylase) is strongly inhibited, whereas the niacin hydroxylase remains unaffected (Fig. 7a).

A new two-step biotransformation in whole cells of *Agrobacterium* DSM 6336 was developed for the preparation of 5-hydroxypyrazinecarboxylic acid (**3** in Fig. 7b) from the starting material 2-cyanopyrazine (**4** in Fig. 7b)⁷⁴. The 5-substituted pyrazine-2-carboxylic acid was then used for the preparation of novel antituberculosis drugs. In previous experiments, various 5-chloropyrazine-2-carboxylic acid esters, pre-

pared from 5-hydroxypyrazinecarboxylic acid, were up to 1,000 times more active than previously described drugs against *Mycobacterium tuberculosis* and other *Mycobacterium* strains⁷⁵.

The biotransformation described involves the hydrolysis of the nitrile group to pyrazinecarboxylic acid (**5** in Fig. 7b) and the subsequent regioselective hydroxylation to 5-hydroxypyrazinecarboxylic acid by the same bacterial cells. The product concentration was as high as 40 g l^{-1} . To induce the required enzymes, the *Agrobacterium* sp. DSM 6336 was first grown on 3-cyanopyridine (**6** in Fig. 7b) as a carbon and energy source, followed by the bioconversion of 2-cyanopyrazine to 5-hydroxypyrazine-2-carboxylic acid (Fig. 7b).

Renewable functionalized pyridines derived from (S)-nicotine

(S)-Nicotine, which is present in concentrations of 2–8% in dried leaves of certain tobacco plants, was used as a starting material for the biocatalytic production of functionalized pyridines from renewable sources. These compounds are precursors in the synthesis of drugs such as analogues of epibatidine (**9** in Fig. 7c), an extremely effective analgesic molecule⁷⁶ that is produced by a tropical frog. The selective functionalization of nicotine at the pyridine or the pyrrolidine ring

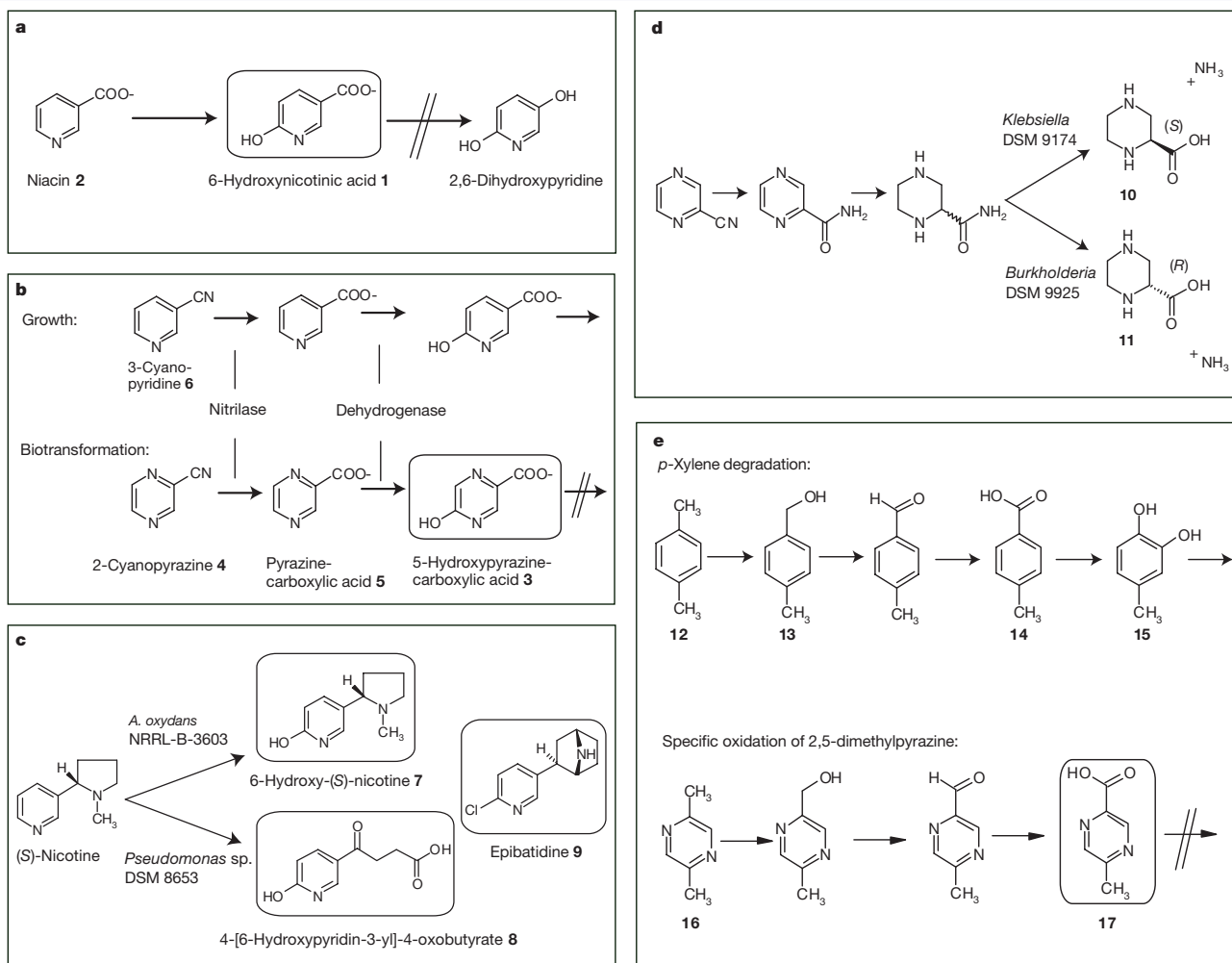


Figure 7 Biocatalytic processes at Lonza. **a**, Catabolic pathway for the production of 6-hydroxynicotinic acid with *Achromobacter xylosoxidans* LK1. **b**, Reaction sequence analogies of 3-cyanopyridine degradation and 2-cyanopyrazine biotransformation with *Agrobacterium* DSM 6336. **c**, Microbial metabolites of (*S*)-nicotine. **d**, Process for the production of enantiomerically pure piperazine-2-carboxylic acids. **e**, Enzymatic oxidation of 2,5-dimethylpyrazine to 5-methylpyrazine-2-carboxylic acid with whole cells of *Pseudomonas putida* ATCC 33015 growing on *p*-xylene as sole carbon source.

by chemical means is difficult to control and this restricts the chemical preparation of semisynthetic pyridines using nicotine as starting material. Although it has long been known that microorganisms form a wide variety of metabolites of nicotine when grown on this heterocycle as the sole carbon and nitrogen source, these metabolites have not been used as starting materials for the chemical synthesis of new nicotine analogues. Using *Arthrobacter oxydans* NRRL-B-3603 and *Pseudomonas* sp. DSM 8653, it was possible to produce up to 30 g l⁻¹ 6-hydroxy-(S)-nicotine (**7** in Fig. 7c) with an isolated yield of 51% and up to 15 g l⁻¹ of 4-[6-hydroxypyridin-3-yl]4-oxobutyrate (**8** in Fig. 7c), respectively⁷⁷.

Preparation of piperazine-2-carboxylic acid

Non-proteinogenic amino acids such as piperazine-2-carboxylic acid are precursors of numerous bioactive compounds. The (S)-analogue (**10** in Fig. 7d) can be used for the synthesis of the HIV protease inhibitor Crixivan from Merck. The best current chemical route to the pure enantiomers of this amino acid is the classical resolution of the racemate by fractional crystallisation of diastereomeric salts. Soil samples containing the racemic carboxamide as the sole nitrogen source were used for enrichments to search for microorganisms with stereoselective amidases. As a result of this screening, *Klebsiella terrigena* DSM 9174 and *Burkholderia* sp. DSM 9925 were used for the preparation of (S)-piperazine-2-carboxylic acid and

(*R*)-piperazine-2-carboxylic acid (**11** in Fig. 7d), respectively. For the production of enantiomerically pure acids the microorganisms were resuspended for various periods of time in buffer containing 10–20 g l⁻¹ of the racemic carboxamide. The enantiomeric excess for both piperazine-2-carboxylic acids was greater than 99% (Fig. 7d)⁷⁸.

Enzymatic oxidation of methyl groups on aromatic heterocycles

Chemical oxidation reactions used for the industrial-scale preparation of heteroaromatic monocarboxylic acids from heteroaromatic compounds bearing one or more methyl groups are often nonspecific and lead to the formation of undesired by-products. To overcome this problem, Lonza has developed a biological oxidation method for this type of reaction.

Wild-type *Pseudomonas putida* ATCC 33015, capable of growing on *p*-xylene as the sole carbon and energy source, was used as the biocatalyst in these investigations. Both the biochemistry and genetics of the xylene degradative pathway have been studied extensively. *p*-Xylene (**12** in Fig. 7e) is oxidized by xylene monooxygenase to 4-methylbenzylalcohol and further oxidized by other enzymes to the aldehyde (**13**) and 4-methylbenzoic acid (**14**). The aromatic carboxylic acid is converted by toluene dioxygenase and a dehydrogenase into 4-methylcatechol (**15**) before cleavage of the aromatic ring by catechol dioxygenase. The cleavage product is then transformed into Krebs-cycle intermediates (Fig. 7e). Lonza has now demonstrated that

P. putida previously grown on *p*-xylene is capable of oxidizing many methylated five- and six-membered rings to the corresponding mono-carboxylic acids, without decarboxylation or ring hydroxylation.

The performance of the enzyme was studied in greater detail on 2,5-dimethylpyrazine (**16** in Fig. 7e). The oxidation product, 5-methylpyrazine-2-carboxylic acid (**17**), is an intermediate for the production of an antilipolytic drug. High product concentrations and high yields were achieved with growing cells. For this reason a mixture of 75% (by volume) *p*-xylene and 25% 2,5-dimethylpyrazine was supplied as growth substrate in large-scale fermentations. After 54 hours, up to 20 g l⁻¹ of 5-methyl-2-pyrazinecarboxylic acid was achieved (Fig. 7e)⁴².

Conversions of steroids, fatty acids and oligosaccharides

A well established industrial application of biocatalysis is the biotransformation of steroids, the products of which are used in contraceptives and other steroid hormone derivatives. Basic building blocks for these important drugs can be derived from natural phytosterols from soya (typically a mixture of β -sitosterol (40%), campesterol (25%) and stigmasterol (25%)), from conifers (tall-oil: mainly β -sitosterol (70%), campesterol (10%) and β -sitostanol (15%)) and from rape seed (β -sitosterol (45%), campesterol (35%) and brassicasterol (12%)). Using mutants of *Mycobacterium* sp. devoid of steroid-ring degradation activities, Schering (Berlin and Bergkamen, Germany) uses these natural raw sources mainly for the production of androstenedione and androsta-dien-dione on a scale of 200 m³. The size of the world market exceeds 1,000 tons per year. These steroids are used as basic substrates for subsequent chemical and biotechnological syntheses of drugs. Biocatalytic processes with economic significance are hydroxylations (for example at the 11 α or 11 β positions with *Curvularia* sp.), dehydrogenations (Δ 1-position; hydrocortisone to prednisolone) and reductions (17-keto-reduction). These processes are run at Schering at a scale below 100 tons per year⁷⁹.

Shimizu⁸⁰ has developed efficient fungal-based systems for the production of polyunsaturated fatty acids (PUFAs). These are precursors of a wide variety of metabolites such as prostaglandins, leukotrienes and hydroxy-fatty acids that regulate critical biological functions. PUFAs are required in every organ for the human body to function normally.

A promising strategy for the application of coenzyme-dependent enzymes in multistep whole-cell biocatalysis has been developed by Kyowa Hakko Kogyo Co. Ltd. (Tokyo, Japan). This approach allows the production of oligosaccharides at high productivity, high product concentration and high yield, and is applied on a cubic-metre scale⁸¹. One product is globotriose, which is produced with a productivity of 5.22 g l⁻¹ h⁻¹ and a final product concentration of 188 g l⁻¹. The concept is based on the application of whole-cell coenzyme regeneration systems and recombinant enzymatic synthesis pathways in different non-viable but metabolically active microorganisms. Reaction intermediates as well as coenzymes diffuse freely among the permeabilized cells.

The next few decades

Cost of whole-cell biocatalytic processes

A detailed analysis of total production costs for the oxidation of *n*-alkanes to terminal *n*-alkanols in a two-phase, whole-cell bioconversion, followed by phase separation and a two-stage product distillation, showed that product costs would be of the order of US\$8 per kg product for both fed-batch and continuous processes, based on a production scale of 10,000 tons per year (ref. 40). This required 800 or 2,000 m³ of bioreactor volume (continuous and batch processes respectively), as well as a processing unit for phase separation, distillation of substrate and product from the carrier solvent, and final separation of product from substrate (also by distillation). The estimated cost distribution was as follows: equipment depreciation and maintenance, 18–22%; substrates and solvents, 20%; energy, 10%; waste handling, 5%; personnel, 2%. The single major

expenditure comprised medium salts and the carbon source (glucose), which accounted for 40% of all costs owing to the limited biocatalyst activities and lifetimes.

A first estimate of whole-cell process feasibility can be based simply on the volumetric activity and catalyst stability that can be expected for a given (recombinant) whole-cell biocatalyst. The rule of thumb is that for products valued at less than US\$20 per kg, the intended production should exceed 1,000 tons per year and volumetric activities should be above 100 international units per litre to warrant further work.

For the present time, biocatalytic processes will therefore undoubtedly be developed for products compatible with costs greater than US\$20–30 per kg. But the above numbers demonstrate that there is significant scope for improvement by increasing biocatalyst activity and stability. This reduces medium costs, as well as the size of the fermentation equipment necessary for given production runs. Synthesis of products with values in the range of US\$5–10 per kg may be feasible, and attempts to develop major products (50,000 tons per year or more) with values of US\$2–5 per kg may be worth considering.

Although research on improved, multiphase bioconversion technology will be useful, it is the biocatalysts that present the most interesting target for improvement. Development of appropriate activities and selectivities enable biocatalytic processes in the first place, and improvements of activity and stability may make a process economically feasible. Thus it is not surprising that research efforts, both for whole-cell systems and isolated enzymes, have focused on improving enzyme properties such as substrate range and specificity, stability and, most interestingly, function in non-classical environments.

Biochips and combinatorial biocatalysis

Nonaqueous enzymology is maturing rapidly. The ability to place virtually any enzyme into an organic solvent and observe catalytic activity is now commonplace. The combination of catalyst engineering with directed evolution and gene-shuffling techniques will yield enzymes with tailored selectivities and synthetically relevant activities in essentially any suitable reaction medium. Increasing the stability of enzymes at solid interfaces and films and coatings⁸² will result in more active and stable immobilized biocatalysts and biocatalytic materials for use in both aqueous and nonaqueous media. More active and stable enzymes at surfaces will enable protein-based chips⁸³ and nanomaterials to be developed for use in proteomics, functional genomics and microfluidic biocatalysis.

Finally, miniaturized biodevices may be combined with combinatorial biocatalysis⁸⁴ to generate libraries of biologically relevant compounds available for rapid screening. The high selectivity of enzymes for chemical-compound library synthesis enables a broad array of natural products and complex synthetics to be derived efficiently using biocatalysts. The mild reaction conditions and high selectivity of biocatalysts are critical for the generation of pharmacologically useful libraries from these lead compounds⁸⁵. Combinatorial biocatalysis goes beyond drug discovery, and many of the same techniques used in pharmaceutical synthesis can be used to generate libraries of agrochemicals, biopolymers, nutraceuticals and separation media.

Continuous regeneration of coenzymes

From a process-engineering perspective, biocatalysts may be coenzyme dependent (such as oxidoreductases) or coenzyme independent (such as hydrolases). Coenzymes such as the pyridine dinucleotides (for example, NAD(P)H) or phosphorylated trinucleotides (such as ATP) are expensive and must be recycled. *In vivo* coenzyme recycling systems have been well established on the laboratory scale during the past two decades⁸⁶. This approach is also used on the industrial scale where it is most easily achieved with intact and metabolically active cells as biocatalysts.

Highly efficient *in vitro* coenzyme recycling methods have been developed and have enabled the use of isolated oxidoreductases on

the laboratory scale (see refs 87 and 88 for reviews). An outstanding application on the industrial scale is the production of *L*-*t*-leucine in a membrane reactor by Degussa with enzymatic recycling of NADH using formate/formate dehydrogenase⁸⁹. Recently, Clair and co-workers (ref. 90) reported a promising new approach for recycling of NADH in a crystalline preparation of horse-liver alcohol dehydrogenase (HLADH-CLEC). Crosslinking of enzyme crystals significantly stabilized the enzyme activity for two days in up to 90% polar organic solvents such as isopropanol. HLADH-CLECs were used in a small packed-bed reactor for the reduction of cinnamaldehyde to cinnamyl alcohol, with butanediol as the source of reduction equivalents and a calculated turnover number for NADH of 12,000.

Electrons can be supplied to oxidoreductases by electrochemical methods. For example, rhodium complexes have been used to channel electrons to NAD(P)H-dependent oxidoreductases, resulting in highly specific reduction of NAD⁺ (refs 91, 92). Mediator proteins have also been used⁷⁹. Attempts to avoid the use of labile and expensive pyridine nucleotides by using artificial redox mediators such as viologens or anthraquinone derivatives have been described for C–H bond-forming oxidoreductases derived from anaerobic microorganisms⁹³. A new *in vitro* regeneration system for ATP has been described recently⁹⁴. Here, the ATP-dependent formation of glucose-6-phosphate from glucose was achieved by coupling polyphosphate:AMP phosphotransferase and adenylate kinase. This allowed the *in situ* formation of ATP from AMP with phosphate groups derived from inexpensive polyphosphate.

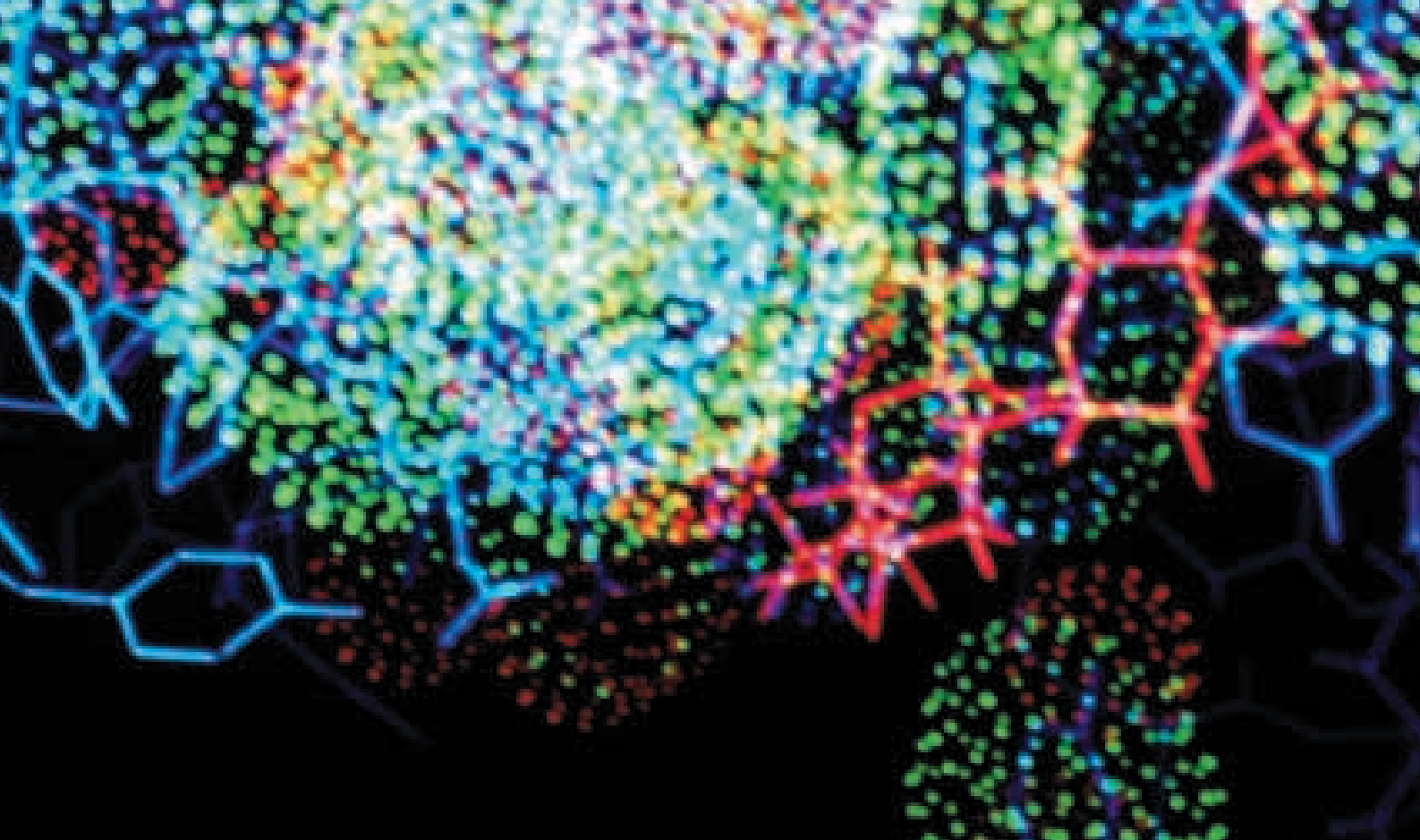
Impact of biocatalysis in the chemical industry

The success of biocatalysis depends ultimately on the economics of specific processes. It provides enormous opportunities; and with the introduction of each new process, experience and confidence accumulate and it becomes easier to develop and justify the next biocatalytic process. This has been the experience at BASF, at DSM, at Lonza, and undoubtedly at other companies that have developed such processes.

Biocatalysis in organic solvents entails many of the same requirements and issues as standard organic chemical processes with respect to equipment, piping and tubing materials, chemicals handling, safety, storage, transportation, GMP and regulatory issues — all of which are already familiar to the chemical industry. The challenge now is to develop a range of useful biocatalysts that can be used in organic synthesis. As this review and the accompanying articles show, the basic tools are abundantly available. There now exist several specialized suppliers of new biocatalytic reagents, such as Novozymes, Genencor, and more recently Diversa and Maxygen. We anticipate that there will be accelerated growth of biocatalytic processes in companies that already have developed some expertise in this area, such as those represented by three of the authors. Other companies currently at an earlier stage in exploiting biocatalysis, such as Dow Chemical and Aventis, will surely expand their activities in this area in the coming decade. □

1. Faber, K. *Biotransformations in Organic Chemistry: A Textbook* (Springer, Berlin, 1997).
2. Dordick, J. S. *Biocatalysts for Industry* (Plenum Press, New York, 1991).
3. Jensen, V. J. & Rugh, S. Industrial scale production and application of immobilized glucose isomerase. *Methods Enzymol.* **136**, 356–370 (1987).
4. Boesten, W. H. J., Moody, H. M. & Roos, E. C. Process for the recovery of ampicillin from enzymatic acylation of 6-aminopenicillanic acid (Chemferm; patent no. WO.9630376; The Netherlands; 1996).
5. Schulze, B., Broxterman, R., Shoemaker, H. & Boesten, W. Review of biocatalysis in the production of chiral fine chemicals. *Spec. Chem.* **18**, 244, 246 (1998).
6. Kaplan, D. L., Dordick, J. S., Gross, R. A. & Swift, G. Enzymes in polymer science: an introduction. *Am. Chem. Soc. Symp. Ser.* **684**, 2–16 (1998).
7. Pokora, A. R. & Cyrus, W. L. Phenolic developer resins (The Mead Corp., Dayton, OH: US patent no. 4,647,952; 1987).
8. Nagasawa, T. & Yamada, H. Application of nitrile converting enzymes for the production of useful compounds. *Pure Appl. Chem.* **62**, 1441–1444 (1990).
9. Buchholz, H. & Kasche, V. *Biokatalysatoren und Enzymtechnologie* (VCH, New York, 1997).
10. Tramper, J. Chemical versus biochemical conversion: when and how to use biocatalysts. *Biotechnol. Bioeng.* **52**, 290–295 (1996).
11. Arnold, F. H. & Volkow, A. A. Directed evolution of biocatalysts. *Curr. Opin. Chem. Biol.* **3**, 54–59 (1999).
12. Stemmer, W. P. C. DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Nature* **370**, 389–391 (1994).
13. Blinkovsky, A. M., Martin, B. D. & Dordick, J. S. Enzymology in monophasic organic media. *Curr. Opin. Biotechnol.* **3**, 124–129 (1992).
14. Klivanov, A. M. Asymmetric transformations catalyzed by enzymes in organic solvents. *Acc. Chem. Res.* **23**, 114–120 (1990).
15. Russell, A. J., Beckman, E. J. & Chaudhary, A. K. Studying enzyme activity in supercritical fluids. *Chemtech* **3**, 33–37 (1994).
16. Lalonde, J., Navia, M. A. & Margolin, A. L. Crosslinked enzyme crystals of lipases as catalysts for kinetic resolution of acids and alcohols. *Methods Enzymol.* **286**, 443–464 (1997).
17. Park, O., Kim, D. Y. & Dordick, J. S. Enzyme-catalyzed synthesis of sugar-containing monomers and linear polymers. *Biotechnol. Bioeng.* **70**, 208–216 (2000).
18. Khmel'nitskiy, Y. L., Welch, S. H., Dlack, D. S. & Dordick, J. S. Salts dramatically enhance activity of enzymes suspended in organic solvents. *J. Am. Chem. Soc.* **116**, 2647–2648 (1994).
19. Ru, M. T., Dordick, J. S., Reimer, J. A. & Clark, D. S. Optimizing the salt-induced activation of enzymes in organic solvents: effects of lyophilization time and water content. *Biotechnol. Bioeng.* **63**, 233–241 (1999).
20. Ru, M. T. *et al.* On the salt-induced activation of lyophilized enzymes in organic solvents: effect of salt kosmotropicity on enzyme activity. *J. Am. Chem. Soc.* **122**, 1565–1571 (2000).
21. Engbersen, J. F. J., Broos, J., Verboom, W. & Reinhoudt, D. N. Effects of crown ethers and small amounts of cosolvent on the activity and enantioselectivity of α -chymotrypsin in organic solvents. *Pure Appl. Chem.* **68**, 2171–2178 (1996).
22. Slade, C. J. & Vulfson, E. N. Induction of catalytic activity in proteins by lyophilization in the presence of a transition state analogue. *Biotechnol. Bioeng.* **57**, 211–215 (1998).
23. Russel, A. J. & Klivanov, A. M. Inhibitor-induced enzyme activation in organic solvents. *J. Biol. Chem.* **263**, 11626–11626 (1988).
24. Khmel'nitskiy, Y. L. *et al.* Synthesis of water soluble Paclitaxel derivatives by enzymatic acylation. *J. Am. Chem. Soc.* **119**, 11554–11555 (1997).
25. Paradkar, V. M. & Dordick, J. S. Aqueous-like activity of α -chymotrypsin dissolved in nearly anhydrous organic solvents. *J. Am. Chem. Soc.* **116**, 5009–5010 (1994).
26. Wang, P., Sergeeva, M. V., Lim, L. & Dordick, J. S. Biocatalytic plastics as active and stable materials for biotransformations. *Nature Biotechnol.* **15**, 789–793 (1997).
27. Yang, Z. *et al.* Activity and stability of enzymes incorporated into acrylic polymers. *J. Am. Chem. Soc.* **117**, 4843–4850 (1995).
28. Chaudhary, A., Lopez, J., Beckman, E. J. & Russell, A. J. Biocatalytic solvent-free polymerization to produce high molecular weight polyesters. *Biotechnol. Prog.* **13**, 318–325 (1997).
29. Wubbolts, M. G., Hoven, J., Melgert, B. & Witholt, B. Efficient production of optically active styrene epoxides in two-liquid phase cultures. *Enzyme Microb. Technol.* **16**, 887–894 (1994).
30. Wubbolts, M. G., Favre-Bulle, O. & Witholt, B. Biosynthesis of synthons in two-liquid phase media. *Biotechnol. Bioeng.* **52**, 301–308 (1996).
31. Held, M. *et al.* Preparative scale production of 3-substituted catechols using a novel monooxygenase from *Pseudomonas azelaica* HBP1. *J. Mol. Catal. Enzymatic* **5**, 87–93 (1998).
32. De Smet, M. J., Wijnberg, H. & Witholt, B. Synthesis of 1,2-epoxycyclohexane by *Pseudomonas oleovorans* during growth in a two phase system containing high concentrations of 1-octene. *Appl. Environ. Microbiol.* **42**, 811–816 (1981).
33. De Smet, M. J., Witholt, B. & Wijnberg, H. Practical approach to high yield enzymatic stereospecific organic synthesis in multiphase systems. *J. Org. Chem.* **46**, 3128–3131 (1981).
34. De Smet, M. J., Kingma, J., Wijnberg, H. & Witholt, B. *Pseudomonas oleovorans* as a tool in bioconversions of organic substrates: growth, morphology and conversion characteristics in different two-phase systems. *Enzyme Microb. Technol.* **5**, 352–360 (1983).
35. Favre-Bulle, O., Schouten, T., Kingma, J. & Witholt, B. Bioconversion of *n*-octane to octanoic acid by a recombinant *Escherichia coli* cultured in a two-liquid phase bioreactor. *Biotechnology* **9**, 367–71 (1991).
36. Laane, C., Boeren, S., Vos, K. & Veeger, C. Rules for optimization of biocatalysts in organic solvents. *Biotechnol. Bioeng.* **30**, 81–87 (1987).
37. Nikolova, P. & Ward, O. P. Whole cell biocatalysis in nonconventional media. *J. Ind. Microbiol.* **12**, 76–86 (1993).
38. Buckland, B. C. *et al.* Microbial conversion of indene to indanol: a key intermediate in the synthesis of Crixivan. *Metabol. Eng.* **1**, 63–74 (1999).
39. Schmid, A., Kollmer, A., Mathys, R. G. & Witholt, B. Developments toward large scale bacterial bioprocesses in the presence of bulk amounts of organic solvents. *Extremophiles* **2**, 249–245 (1998).
40. Mathys, R. G., Schmid, A. & Witholt, B. Integrated two-liquid phase bioconversion and product-recovery processes for the oxidation of alkanes: process design and economic evaluation. *Biotechnol. Bioeng.* **64**, 459–477 (1999).
41. Panke, S., Witholt, B., Schmid, A. & Wubbolts, M. G. Towards a biocatalyst for (S)-styrene oxide production: characterization of the styrene degradation pathway of *Pseudomonas* sp. strain VLB120. *Appl. Environ. Microbiol.* **64**, 2032–2043 (1998). [Published erratum appears in *Appl. Environ. Microbiol.* **64**, 3546 (1998).]
42. Kiener, A. Enzymatic oxidation of methyl groups on aromatic heterocycles: a versatile method for the preparation of heteroaromatic carboxylic acids. *Angew. Chem. Int. Ed. Engl.* **31**, 774–775 (1992). (Patent no. EP-B 0442430.)
43. Lamare, S. & Legoy, M.-D. Biocatalysis in the gas phase. *Trends Biotechnol.* **11**, 413–418 (1993).
44. Held, M. *et al.* An integrated process for the production of toxic catechols from toxic phenols based on a designer biocatalyst. *Biotechnol. Bioeng.* **62**, 641–648 (1999).
45. Li, Z., Feiten, H.-J., Van Beilen, J. B., Duetz, W. & Witholt, B. Preparation of optically active *N*-benzyl-3-hydroxy-pyrrolidine by enzymatic hydroxylation. *Tetrahedr. Asym.* **10**, 1323–1333 (1999).
46. Schmid, A., Sonnleitner, B. & Witholt, B. Medium chain length alkane solvent-cell transfer rates in two-liquid phase, *Pseudomonas oleovorans* cultures. *Biotechnol. Bioeng.* **60**, 10–23 (1998).
47. Schmid, A., Kollmer, A., Mathys, R. G. & Witholt, B. Developments toward large-scale bacterial bioprocesses in the presence of bulk amounts of organic solvents. *Extremophiles* **2**, 249–256 (1998).
48. Panke, S., Wubbolts, M. G., Schmid, A. & Witholt, B. Production of enantiopure styrene oxide by recombinant *Escherichia coli* synthesizing a two-component styrene monooxygenase. *Biotechnol. Bioeng.* **69**, 91–100 (2000).
49. Panke, S., de Lorenzo, V., Kaiser, A., Witholt, B. & Wubbolts, M. G. Engineering of a stable whole-cell biocatalyst capable of (S)-styrene oxide formation for continuous two-liquid-phase applications. *Appl. Environ. Microbiol.* **65**, 5619–5623 (1999).
50. Yamada, H. in *New Frontiers in Screening for Microbial Biocatalysts* (eds Kieslich, K., van der Beek, C. P., de Bont, J. A. M. & van den Tweel, W. J. I.) 13–17 (Elsevier, Amsterdam, 1998).
51. Lye, G. J. & Woodley, J. M. Application of *in situ* product-removal techniques to biocatalytic processes. *Trends Biotechnol.* **17**, 395–402 (1999).

52. Witholt, B. *et al.* Bioconversions of aliphatic compounds by *Pseudomonas oleovorans* in multiphase bioreactors: background and economic potential. *Trends Biotechnol.* **8**, 46–52 (1990).
53. Mathys, R. G., Schmid, A., Kut, O. M. & Witholt, B. Alkanol removal from the apolar phase of a two-liquid fermentation system. 2. Effect of fermentation medium on distillation. *J. Chem. Technol. Biotechnol.* **71**, 326–334 (1998).
54. Mathys, R. G., Kut, O. M. & Witholt, B. Alkanol removal from the apolar phase of a two-liquid fermentation system. 1. Comparison of a less volatile and more volatile *in situ* extraction solvent for separation of 1-octanol by distillation. *J. Chem. Technol. Biotechnol.* **71**, 315–325 (1998).
55. Balkenhohl, F., Dittrich, K., Hauer, B. & Ladner, W. Optically active amines via lipase-catalyzed methoxyacetylation. *J. prakt. Chemie* **339**, 381–384 (1997).
56. Hauer, B. *et al.* The development of enzymes for the preparation of chemicals. *Chimia* **53**, 613–616 (1999).
57. Ress-Löschke, M., Friedrich, T., Hauer, B., Mattes, R. & Engels, D. Verfahren zur Herstellung chiraler Carbonsäuren (BASF; German patent no. DE 198 448 129 A1; 1998).
58. Boesten, W. H. J. Enzymic Division of DL-phenylglycinamide into its optically active antipodes (German patent no. DE 2526594; 1976).
59. Boesten, W. H. J. & Meyer-Hoffman, L. R. M. Enzyme product with L- α -aminoacylamidase activity (NOVO, GB 1577087; 1977).
60. Broxterman, R., Sonke, T., Worries, H. & van den Tweel, W. Biocatalytic production of unnatural amino acids. *Pharm. Manuf. Int.* **61**(2000).
61. Broxterman, Q. B., Boesten, W. H. J., Schoemaker, H. E. & Schulze, B. DSM gets the biocatalysis bug. *Spec. Chem.* **17**, 186–193 (1997).
62. Boesten, W. H. J., Broxterman, Q. B. & Plaum, M. J. M. Process for preparing optically active 2-amino- α -oxoalkanoic acid derivatives (DSM, EP 905257; 1998).
63. Kaptein, B. *et al.* Preparation and use of enantiomerically pure C α -tetrasubstituted α -amino acids. *Chim. Oggi* **14**, 9–12 (1996).
64. Kaptein, B. *et al.* Enzymic resolution of α,α -disubstituted α -amino acid esters and amides. *Tetrahedr. Asym.* **4**, 1113–1116 (1993).
65. Kruizinga, W. H. *et al.* Synthesis of optically pure α -alkylated α -amino acids and a single-step method for enantiomeric excess determination. *J. Org. Chem.* **53**, 5392 (1988).
66. Roos, E. C. *et al.* Synthesis of racemic α -amino carboxamides via Lewis acid-mediated reactions of α -methoxyglycinamide derivatives with allylsilanes: enzymic resolution to optically active α -amino acids. *J. Org. Chem.* **57**, 6769–6778 (1992).
67. Schoemaker, H. E. *et al.* Application of enzymes in industrial organic synthesis. *Chimia* **51**, 308–310 (1997).
68. Kirchner, G., Salzbrenner, E., Werenka, C. & Boesten, W. H. J. Process for the preparation of the disodium salt of Z-L-aspartic acid from fumaric acid (DSM Chemie and Holland Sweetner Company, EP 832982; 1997).
69. Bruggink, A. Biocatalysis and process integration in the synthesis of semisynthetic antibiotics. *Chimia* **50**, 431–432 (1996).
70. Demain, A. L. & Baez-Vasquez, M. A. Immobilized *Streptomyces clavuligerus* NP1 cells for biotransformation of penicillin G into deacetoxycephalosporin G. *Appl. Biochem. Biotechnol.* **87**, 135–140 (2000).
71. Boesten, W. H. J., van Dooren, T. J. & Smeets, J. C. M. Process for the preparation of a β -lactam antibiotic (Chemferm, WO 9623897; 1996).
72. Bruggink, A., Roos, E. C. & de Vroom, E. Penicillin acylase in the industrial production of β -lactam antibiotics. *Org. Process Res. Dev.* **2**, 128–133 (1998).
73. Kulla, H. G. Enzymatic hydroxylations in industrial application. *Chimia* **45**, 81–85 (1991). (Lonza, EP-B 0152948; EP-B 0152949.)
74. Wieser, M., Heinzmann, K. & Kiener, A. Bioconversion of 2-cyanopyrazine to 5-hydroxypyrazine-2-carboxylic acid with *Agrobacterium* sp. DSM 6336. *Appl. Microbiol. Biotechnol.* **48**, 174–176 (1997). (Lonza, EP-B 0578137.)
75. Bergmann, K. E., Cynamon, M. H. & Welch, J. T. Quantitative structure-activity relationships for the in-vitro antimycobacterial activity for pyrazinoic acid esters. *J. Med. Chem.* **39**, 3394–3400 (1996).
76. Spande, T. F. *et al.* Epibatidine: A novel (chloropyridyl) azabicycloheptane with potent analgesic activity from an Ecuadorian poison frog. *J. Am. Chem. Soc.* **114**, 3475–3478 (1992).
77. Roduit, J.-P., Wellig, A. & Kiener, A. Renewable functionalized pyridines derived from microbial metabolites of the alkaloid (S)-nicotine. *Heterocycles* **46**, 1687–1702 (1997). (Lonza, EP-B 0713869.)
78. Eichhorn, E., Roduit, J.-P., Shaw, N., Heinzmann, K. & Kiener, A. Preparation of (S)-piperazine-2-carboxylic acid, (R)-piperazine-2-carboxylic acid, and (S)-piperidine-2-carboxylic acid by kinetic resolution of the corresponding racemic carboxamides with stereoselective amidases in whole bacterial cells. *Tetrahedr. Asym.* **8**, 2533–2536 (1997). (Lonza, EP-A 0686698; WO 96/35775.)
79. Reipa, V., Mayhew, M. P. & Vilker, V. L. A direct electrode driven P450 cycle for biocatalysis. *Proc. Natl Acad. Sci. USA* **94**, 13554–13558 (1997).
80. Certik, M. & Shimizu, S. Biosynthesis and regulation of microbial polyunsaturated fatty acid production. *J. Biosci. Bioeng.* **87**, 1–14 (1999).
81. Koizumi, S., Endo, T., Tabata, K. & Ozaki, A. Large scale production of UDP-galactose and globotriose by coupling metabolically engineered bacteria. *Nature Biotechnol.* **16**, 847–850 (1998).
82. Kim, Y., Dordick, J. S. & Clark, D. S. *Biotechnol. Bioeng.* (in the press).
83. MacBeath, G. & Schreiber, S. L. Printing proteins as microarrays for high-throughput function determination. *Science* **289**, 1760–1763 (2000).
84. Michels, P. C., Khamelnitsky, Y. L., Dordick, J. S. & Clard, D. S. Combinatorial biocatalysis: a natural approach to drug discovery. *Trends Biotechnol.* **16**, 210–215 (1998).
85. Krstenansky, J. L. & Khamelnitsky, Y. L. Biocatalytic combinatorial synthesis. *Bioorg. Med. Chem.* **7**, 2157–2162 (1999).
86. Simon, H., Bader, J., Günther, H., Neumann, S. & Thanos, J. Chiral compounds synthesized by biocatalytic reductions. *Angew. Chem. Int. Ed.* **24**, 539–553 (1985).
87. Adlercreutz, P. Cofactor regeneration in biocatalysis in organic media. *Biocatal. Biotransform.* **14**, 1–30 (1996).
88. Chenault, H. K., Simon, E. S. & Whitesides, G. M. Cofactor regeneration for enzyme catalysed synthesis. *Biotechnol. Gen. Eng. Rev.* **6**, 221–270 (1988).
89. Bommarius, A. S., Drauz, K., Groeger, U. & Wandrey, C. in *Chirality in Industry* (eds Collins, A. N., Sheldrake, G. N. & Crosby, J.) 371–397 (Wiley, Chichester, 1992).
90. Clair, N. S., Wang, Y.-F. & Margolin, A. L. Cofactor-bound cross linked enzyme crystals (CLEC) of alcohol dehydrogenase. *Angew. Chem. Int. Ed.* **39**, 380–383 (2000).
91. Westerhausen, D., Herrmann, S., Hummel, W. & Steckhan, E. Formate driven, non-enzymatic NAD(P)H regeneration for the alcohol dehydrogenase catalyzed stereoselective reduction of 4-phenylbutanone. *Angew. Chem. Int. Ed.* **31**, 1529–1531 (1992).
92. Hollmann, F., Schmid, A. & Steckhan, E. First synthetic application of a monooxygenase employing indirect electrochemical NADH-regeneration. *Angew. Chem. Int. Ed.* (in the press).
93. Günther, H. & Simon, H. Artificial electron carriers for preparative biocatalytic redox reactions forming reversibly carbon hydrogen bonds with enzymes present in strict or facultative anaerobes. *Biocatal. Biotransform.* **12**, 1–26 (1995).
94. Resnick, S. & Zehnder, A. J. B. *In vitro* ATP regeneration from polyphosphate and AMP by polyphosphate:AMP phosphotransferase and adenylate kinase from *Acinetobacter johnsonii* 210A. *Appl. Environ. Microbiol.* **66**, 2045–2051 (2000).



NOVOZYMES DRIVEN BY RESEARCH AND SCIENTISTS

For 50 years we have consistently demonstrated a deep-rooted commitment to research. By setting the agenda for the industrial application of microbiology and protein chemistry, we have been pioneers in biotechnology and enzymology.

Today, Novozymes is the world's largest manufacturer of industrial enzymes, with over 500 products. We have more than 3,000 employees and close ties to leading research institutions throughout the world.

Research strategy Novozymes' research goal is to build on our leading technological platform in each of the integral biotechnological disciplines needed for the production of proteins. From the identification and development of unique enzymes and viable new enzyme

applications, to the efficient expression and manufacture of the products, our research expertise provides the infrastructure to build technological expertise from the ground up – from the initial phases of research through to the product launch.

Formerly part of Novo Nordisk, Novozymes has amassed a collection of 25,000 classified fungal and bacterial cultures. Our strength lies in finding new products by drawing not only on this collection and modern techniques of recombinant DNA and directed molecular evolution, but also on an international network of microbiology centres that supplement and extend our access to Nature in all her diversity.

We have a portfolio of more than 4,000 active patents, patent applications and licensed patents.

Technologies Our business builds on the following technologies:

- Microbiology: assays, screening and mutagenesis, all in an efficient automated set-up
- Bioinformatics: DNA sequencing, in silico biology, DNA microarrays and functional genomics
- Recombinant DNA expression, gene technology: a wide variety of novel bacterial and fungal hosts and tools to ensure efficient yields of secreted proteins
- Protein chemistry: protein structure/-function, mechanistic studies, biochemical characterization and purification from micro scale to large scale
- Directed molecular evolution: site directed and random mutagenesis and gene shuffling for greater diversity generation
- Screening: automated high-throughput facilities for very large libraries



Unlocking the magic of nature

Novozymes – scientific firsts

Novozymes is the world's largest discoverer, manufacturer and marketer of industrial enzymes, with over 500 products. We have more than 3,000 employees and close ties to leading research institutions throughout the world. Our R&D organization is home to 650 highly trained employees, including 280 graduate-level scientists.

With few exceptions we have been the first to launch new enzyme applications and widen the market for industrial enzymes. For example, we introduced proteases and lipases for industrial use, amylases for sweeteners, as well as the industrial application of cellulases and oxidoreductases.

1941	Trypsin, the first extracted protease
1952	Thermozyme® (BAN), the first fermented enzyme from Novozymes
1963	Alcalase®, the protease that revolutionized the market for industrial enzymes and established the use of enzymes in the detergent industry
1973	Termamyl®, a thermostable alpha-amylase for the starch industry
1974	Esperase®, the first extremophilic protease, stable up to a pH of 12
1974	Sweetzyme®, the first glucose isomerase
1984	Maltogenase®, a maltogenic amylase for the starch industry, the first enzyme produced using genetically modified microorganisms
1987	Celluzyme®, a multicomponent enzyme, the first alkaline cellulase for the detergent industry
1988	Denimax®, the first alkaline cellulase for stone-washing denim
1988	Lipolase®, the first lipolytic detergent enzyme and the first detergent enzyme produced using gene technology
1991	Durazym®, a bleach-stable detergent protease developed using computer simulation
1991	Carezyme®, the first monocomponent alkaline cellulase for detergents
1996	DeniLite®, the first industrial laccase and the first bleaching enzyme acting with the help of a mediator molecule
1997	Kannase®, a low-temperature protease produced using directed evolution
1998	Lipoprime™, the first detergent lipase developed using molecular evolution
1999	BioPrep™, a pectate lyase for the biopreparation of cotton
1999	The first microbial peroxidase for bleaching dyes in solution
2000	Mannaway™, the first mannanase enzyme for in-wash stain removal



Today our industrial enzymes can improve the strengths of nature, so we can reduce waste and replace harmful substances.

- Fermentation and recovery: laboratory, pilot and full scale
- Formulation: advanced technologies for both solid and liquid products, including zero dust
- Reduced allergenicity: advanced methods for chemically modifying proteins and determining critical sites for protein-engineering reduced allergenicity into the protein structure

Knowledge and partnerships Focusing on knowledge and partnerships is integral to our business. Examples include broad-based alliances with biotechnology centres throughout the world and cooperation with companies like Maxygen Inc. on gene shuffling and molecular breeding, and Enzon Inc. on minimizing enzyme-induced allergic reactions.

Our pursuit of mutually rewarding partnerships extends further – we also apply our knowledge in close association with customers to develop biotech products.

We are always on the lookout for new partnerships that enable us to use our underlying expertise in novel ways. This means that Novozymes is also keen to employ technologies outside our traditional core business of industrial enzymes.

First with new enzymes Novozymes created the industrial enzyme market. In the last five years alone Novozymes has launched 31 new products, 19 of them containing new enzyme molecules, and we expect to launch an additional 15-25 new products by 2002.

With few exceptions we have been the first to launch new enzyme applications and widen the market for industrial enzymes.

Worldwide Our customers include major players in a wide variety of industries in numerous global markets, the most dominant being the detergent industry. Other important customers include the starch, textile, leather, paper, baking, brewing, wine, juice, alcohol, food and feed industries.

Novozymes has established an impressive worldwide service network together with research centres in Denmark, the USA, China and Japan and production facilities in Denmark, China, Brazil, Switzerland and the USA.



Unlocking the magic of nature